DOCUMENT RESUME

ED 238 935 TM 840 037

AUTHOR	Herman, Joan; Webb, Noreen
TITLE	Item Structures for Diagnostic Testing. Methodology Project.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	30 Nov 83
GRANT	NIE-G-83-0001
NOTE	119p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Diagnostic Tests; *Educational Improvement; Elementary Secondary Education; *Evaluation Methods; Language Arts; Models; School Districts; Sciences; *Student Problems; *Test Construction; Test Items; Test Results; Test Use

ABSTRACT

This paper describes a four-step approach to constructing a diagnostic test that provides precise but practical information on students' problems and needs for additional instruction or remediation. The approach is based on analyzing the structure of the domain to determine which skills within the domain need to be assessed to diagnose students' problems. The four steps include: (1) identifying the factors that describe the curricular domain, (2) constructing a test with items representing all possible combinations of content and cognitive factors, (3) determining which factors and interactions among them produced variations in students' scores using generalizability theory, and (4) determining the minimum number of items needed to obtain a generalizable measure of each skill in the diagnostic profile. This paper contains three studies: "Optimizing the Diagnostic Power of Tests: An Illustration from Language Arts," by Noreen Webb, Joan Herman, and Beverly Cabello; "Diagnosing Student Errors: An Example from Science," by Steve Shaha; and, "Task Structure Design: Beyond Linkage," by Eva Baker and Joan Herman. Two of these studies explicity address problems related to the design of diagnostic tests; the third considers a conceptual model for integrating testing and instruction. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED238935

METHODOLOGY PROJECT
DELIVERABLE - November 30, 1983

ITEM STRUCTURES FOR DIAGNOSTIC TESTING

by

Joan Herman and Noreen Webb
Study Directors

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Grant Number
NIE-G-83-0001

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

G. Grey

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TM 840 037

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

PREFACE

CSE's Methodology Project during FY1983 had two primary emphasis: diagnostic testing and comprehensive evaluation systems for local school improvement. Within the former area, CSE conducted a series of research activities to investigate the feasibility and potential of diagnostic testing for classroom use. Preparatory to the development of a revised microcomputer-aided diagnostic testing system, the research addressed issues in test design and analysis and in the potential and applications of alternative models of diagnosis. The results of each of these inquiries is reported separately.

This document reports on CSE's research efforts in the area of test design. Of the three studies included here, two explicitly address problems related to the design of diagnostic tests; the third considers a conceptual model for integrating testing and instruction.

The first paper, "Optimizing the Diagnostic Power of Tests: An Illustration from Language Arts," investigates strategies for improving the diagnostic power of test items so that they provide more precise but practical information on students' problems and needs. Based on a domain-referenced approach, the study examines factors which may be diagnostically useful in profiling students' performance and explores methods for analyzing and structuring diagnostic tests.

The second paper, "Diagnosing Student Errors: An Example from Science," investigates the effects of cognitive level on student test performance and examines the utility of particular error types in characterizing student strengths and weaknesses.

The final paper, "Task Structure Design: Beyond Linkage" presents a conceptual model for designing testing to maximize the integration of testing and instruction and to maximize the utility of test results for instructional decisionmaking.

Table of Contents

Paper 1 Optimizing the Diagnostic Power of Tests:
An Illustration from Language Arts
By Noreen Webb, Joan Herman, and Beverly Cabello

Paper 2 Diagnosing Student Errors:
An Example from Science
By Steve Shaha

Paper 3 Task Structure Design:
Beyond Linkage
By Eva L. Baker and Joan L. Herman

OPTIMIZING THE DIAGNOSTIC POWER OF TESTS: AN ILLUSTRATION FROM LANGUAGE ARTS

by

Noreen Webb, Joan Herman and Beverly Cabello

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

		Page
Part I:	Introduction	1
Part II:	A Domain Referenced Approach to Test Design	3
Part III:	Development and Administration of the Test	6
	1. Design of the Test	6
	2. Test Administration	12
Part IV:	Overview of the Analytic Approach	13
	1. Traditional Approach to Internal Consistency	13
	2. The Analytic Questions	14
	3. Multidimensional Approach to Test Structure: Generalizability Theory	15
Part V:	Results of Illustrative Analyses	24
	1. Preliminary Analyses	24
	2. Summary of the Three Designs	25
	3. Variance Components and Descriptive Analyses	25
	4. Primary Sources of Variation and Example Diagnostic Profiles	34
	5. The Optimal Number of Items	38
Part VI:	Distractor Analysis	41
	1. Method	42
	2. Results	42
Conclusions		46
References		50

List of Tables & Figures

Tables

Page

Table 1 Design of the Pronoun Test 10

Table 2 Proportion of Total Variation Accounted for by Each Variance Component 27

Table 3 Descriptive Results for Major Sources of Variation in Design I 29

Table 4 Descriptive Results for Major Sources of Variation in Design II 30

Table 5 Descriptive Results for Major Sources of Variation in Design III 31

Table 6 Number of Items Corresponding to Different Generalizability Coefficients 39

Table 7 Agreement on distractor choices for parallel pairs of items 43

Figures

Figure 1 Individual Profiles for Three Students 35

Figure 2 Mean Profiles for FEP and LEP Students 37

Introduction

Assessment has an integral role to play in the improvement of instructional practice. Mastery learning strategies (Bloom, 1976; Block, 1971), systematic instruction (Popham and Baker, 1976), individualized instruction (Glaser, 1970; Klausmeier 1976); clinical teaching (Hunter, 1983) and effective schooling (Edmonds, 1981) all point to the importance of assessment in diagnosing students' strengths and weaknesses, in monitoring their progress through the curriculum, in providing instruction that is tailored to instructional needs and goals and thus in enhancing student achievement. The underlying theory derives from a systems view of education and suggests that if teachers are to maximize their students' learning, they need to: plan instruction on the basis of the needs of individual or groups of students; to monitor their progress; to determine whether remediation is required; and to evaluate outcomes to assess the success of instruction, as well as needs for modification and students' readiness for succeeding work.

Diagnosis and prescription is thus a recurring concern throughout the instructional process and is central to its success. Yet despite its importance, the assessment tools teachers have available for such a process are really quite limited. While so-called diagnostic tests do exist, the level of information they provide is less than optimal. A typical diagnostic test in reading, for example, may characterize student needs by providing a total score and subscores for individuals and groups in such areas as vocabulary, literal comprehension, inferential comprehension, etc, but such scores offer teachers little guidance regarding the nature of any reading problems or their causes. It is left to the teacher to pinpoint why students perform as they do and to prescribe instruction accordingly. In contrast to this global approach, more recent research has taken a molecular view of the diagnostic problem. Tatsuoka and associates (1980) for example, have completed extensive work in diagnosing student performance in a very narrow mathematics domain (the subtraction of two digit signed numbers) and have identified the specific misconceptions and difficulties which students manifest in this area, e.g., six specific error types related to determining the sign of answers. While the advent of classroom computer technology may make such advances more useable in future classroom practice, these findings provide a level of detail beyond the grasp of today's teachers: a teacher cannot track a classroom of students across so many error dimensions for all curriculum areas, nor feasibly tailor instruction at this level of specificity.

The current study seeks an intermediate level for constructing and analyzing diagnostic tests for classroom use. It investigates strategies for improving the power of diagnostic instruments so that they provide more precise but practical information on students' problems and needs. Based on a domain-referenced approach (Hively et al, 1973; Baker, 1974; Popham 1980), the study examines factors which may be diagnostically useful in characterizing or profiling students performance across a range of content areas or domains i.e., factors which may be used to structure the test domain, which predict and conceptually define item difficulty, and which likewise may be used to structure instructional treatments. The study also explores methods for analyzing and structuring diagnostic tests so that the process is efficient and conservative of the information load on teachers. Specifically the study addresses four inter-related questions:

1. What factors ought to be considered in specifying a test domain so that the resultant test will provide specific instructionally relevant diagnostic information?
2. What analyses procedures can be used to optimally structure a diagnostic test to provide valid, reliable, and efficient profiles of individual and group performance?
3. What additional information can be gained from an analysis of students' wrong answer choices? Can the diagnostic value and efficiency of the resultant test be increased?
4. Are the subject strategies feasible for classroom use, or do they require a unrealistic investment in time or an impractical level of detail?

In order to address these questions, we developed an illustrative test of pronoun use representing factors of interest; administered the test to a heterogeneous sample of sixth grade students; used generalizability theory to analyze results and suggest an optimal test structure; examined the consistency and implications of students' distractor choices; and finally reflected on the entire process to assess feasibility and implications for practice. In the sections which follow, we first describe the domain referenced framework which guided the test development process and the specific factors which were chosen for scrutiny, followed by a description of the test, the analytical approach, and the results of our inquiry.

A Domain Referenced Approach to Test Design

A domain referenced approach to test design starts with the assumption that the major purpose of testing is to assess an individual's status with respect to a skill or knowledge domain and further that valid assessment of that status requires a thorough understanding and specification of the domain to be assessed. The objective of assessment, in other words, needs to be well defined to assure that a test actually measures what it is intended to measure and that items reflect test content. The definition is reflected in a domain specification which provides a blueprint for developing test items and can serve also to target effective instructional sequences.

While a number of approaches to domain specification have been proposed, all seek to define a pool of items that represents an important universe of knowledge or skill domain such that student performance in one set of items drawn from the domain would generalize to a second set of items and to the entire domain. In its most highly prescribed form, domain specifications provide an exhaustive set of rules for generating a set of test items (Hively et al, 1973; Osburn, 1968; Millman, 1980). As more commonly practiced (and as exemplified in the present study) domain specifications provide a conceptual map of the skill to be assessed, including relevant parameters for defining the range of eligible content, the response level to be represented in the item, item format, directions, and a sample item (Baker; 1974, Popham, 1980; Hambleton, 1980).

Regardless of approach, the identification of relevant parameters becomes a central problem. Establishing content limits is an initial concern, most commonly solved by reference to extant curricular material, subject area specialists and/or mutually agreed upon goals and boundaries, -- or more preferably research on the structure of the knowledge base and the nature of learning and development. Establishing response limits, including criteria for judging constructed responses and rules for generating incorrect alternatives in selected responses, fixes attention on the quality of expected performance, the level of response differentiation desired and systematic error patterns that may be operable. Framed by linguistic complexity, form of content, and cognitive complexity, test content is specified to represent the domain of interest (see Baker and Herman, 1983).

In addition to identifying content parameters which must be included to assure that a test provides a representative picture of a skill, diagnostic tests present the additional problem of isolating factors which influence variations in student performance and predict varying levels of skill proficiency. In other words, what important factors within a domain cause an item to be more or less difficult or a student's performance to vary. Items representing these factors then can be appropriately sampled to produce a test with diagnostic utility, i.e., one which identifies the causes or reasons for students performance level.

What variables might be useful for constructing such diagnostic profiles? Research in cognitive psychology provides some clues. Chi and Glaser (1980) propose a framework for understanding the nature of differences between expert and novice performance. Their framework characterizes information processing in terms of two components: knowledge or content structure, and cognitive processes, components which are well supported in the research literature. Various authors, for example have pointed to the effect of cognitive processing demands inherent in a task. Principal distinctions have been made for tasks which require storage, association, and retrieval of information contrasted with tasks requiring processing of information, including subordination, reconfiguration and other adaptive processes (Spiro, 1980; Quellmalz, 1982).

Beyond their theoretical justification, content structure and cognitive complexity are appealing also in terms of their feasibility for practical use. Teachers of course are well used to dealing with the structure of content (-- at least as their curriculum or instructional materials define it), and their coverage of that content, as research (and intuitive logic) amply demonstrate is strongly related to student test performance. Cognitive complexity, while perhaps not in the common parlance of classroom teachers, can be operationally defined to be easily accessible to them.

The present study is derived from the foregoing framework of domain referenced testing. A diagnostic test was developed to assess one skill within the language arts curriculum. The domain and item pool were developed to assess the effects on student performance of content structure and cognitive complexity to examine their utility for constructing diagnostic profiles. The test development process is described in the section which follows.

Development and Administration of the Test

Design of the Test

After selecting language arts as a target area for test development, local teachers and administrators were asked to indicate the kinds of grammar problems their students most frequently exhibited at the upper elementary and junior high school grade levels. One of the most common responses was that students have difficulty with pronouns particularly in identifying the correct pronoun referent or in using pronouns correctly. They also indicated that a diagnostic test of pronoun use would be beneficial for their classroom instruction. Pronoun use was therefore selected as an appropriate topic for diagnostic test development.

Following the procedures outlined above, language curricula, texts and content experts were consulted to specify the test domain. Specifications of linguistic properties, e.g. the recurrence and complexity and sequencing of the vocabulary and phrases were also included to assure that the language would be clear and comprehensible to the test taker and that the test could therefore be a measure pronoun use rather than reading comprehension, (see, for example, Doehring and Aulls, 1979). Distractor rules were developed systematically to reflect common usage errors. The domain specification reflected in particular the two factors selected for inquiry.

The content structure factor. The curricular review showed that nominative, objective, (including direct object, indirect object and object of the preposition) and possessive pronouns appear most frequently. These five types of pronouns (including the three objective forms) correspond to rules of grammar, and are called pronoun rules in this paper. The review further revealed that the pronouns corresponding to each rule can also be classified by form.

number and person. There are two types of form: relative form (who or whom) and non-relative form. Number pertains to singular (she) and plural (they). Person can be of three types: first (I, we), second (you), and third (he, they). Since items measuring the second person would have sounded contrived to the reader, the test developed here included only the first and third persons.

The cognitive complexity factor. The two levels of cognitive complexity corresponded to whether students had to use the context of a reading passage to determine the correct pronoun. In the first level, the student was presented with a single sentence that included an underlined noun(s). The student was to select the pronoun to match the underline noun(s). In other words, the pronoun referent was given and the student need only associate that referent with the correct pronoun. In the second, more complex level, students were presented with a short paragraph that included a blank in the place of one noun; students needed to use the context of the paragraph to identify the referent that was appropriate to the blank and then select the correct pronoun for that referent. The correct pronoun could be determined only from elements of the paragraph in which the pronoun was embedded. Consequently, the test developed here used two levels of embeddedness corresponding to two levels of cognitive complexity; non-embedded items (a single sentence) and embedded items (a paragraph).

In summary, the test had five pronoun factors; including four representing content structure and one representing cognitive complexity: pronoun rule (nominative, three types of objective, possessive), pronoun form (relative, non-relative), pronoun number (singular, plural), pronoun person (first, third), and embeddedness of the pronoun (single sentence, paragraph).

Structure of the test. To investigate the impact of each factor on test performance, items were generated for as many combinations of the factors as possible. For each combination, two parallel items were written. The ideal test would have items for every combination of the five factors. Since the form, embeddedness, person, and factors each had two levels and the rule factor had five levels, a complete test would have 80 (2 X 5) combinations. However, for several combinations of factors, sensible items could not be written. First, non-embedded items could not be written to elicit singular first person pronouns (I, me, or my). Second, items testing the relative form of first-person pronouns would have been contrived. Third, there exist no relative form of possessive pronouns. Excluding these combinations of factors leaves 46 combinations. Since two parallel items were written for each combination, the total test had 92 items. The total design of the test is presented in Table 1.

The analytic approach used here to analyze the test structure requires a fully crossed, balance design. Since the design of the total test was unbalanced--34 cells in Table 1 are empty--it was necessary to divide the total design into three fully crossed, balanced designs to represent all cells in the design. Design I represented the combination of five factors: for (2 levels), embeddedness (2 levels), rule (4 levels), number (2 levels), and items (2 levels). This design had 64 items. As indicated in Table 1, the

Table 1
Design of the Pronoun Test

	Non Relative Pronoun								Relative Pronoun							
	Non-Embedded				Embedded				Non-Embedded				Embedded			
	1st Person		3rd Person		1st Person		3rd Person		1st Person		3rd Person		1st Person		3rd Person	
Rule:	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.
Nominative	—[a]	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Direct Object	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Indirect Object	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Object of Preposition	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Possessive	—	2	2	2	—	2	2	2	—	—	—	—	—	—	—	—

[a] No items in this cell.
[b] 2 items per cell.

inclusion of the form factor made it impossible to include items measuring first person pronouns and items measuring possessive pronouns.

The two remaining designs were formed to include the possessive rule. Since the possessive rule applies only to non-relative pronouns, these two designs consisted only of non-relative items. One design (Design II) incorporated the contrast between singular and plural pronouns (number). The other design (Design III) incorporated the constrast between first person and third person pronouns (person). Design II, then, included four factors: embeddedness (2 levels), rule (5 levels), number (2 levels), and item (2 levels), resulting in 40 items. Design III also included four factors: embeddedness (2 levels), rule (5 levels), person (2 levels), and item (2 levels), resulting in 40 items. Many items in the test were included in more than one of the three designs. All of the analyses presented in this paper focus on these three designs.

Structure of the item. The test used a multiple choice format with five responses per item. Three distractors were correct in all ways but one. The fourth distractor was correct in only one way or not at all. An example is the following item, "Mom praised Mary and Stevie", with the following responses: them, they, us, him and she. The correct response (them) is an objective, plural third-person pronoun. The next three responses (they, us, and him) were correct on two of the three factors (rule, number, person). The final response (she) was correct only in the person. The last response was considered a "wild card" distractor (a highly unlikely selection).

Such distractors were included to detect guessing or carelessness.

Test Administration

Through pilot administrations and feedback from teachers and students, the test was modified three times. The final diagnostic test was administered to 128 sixth-grade students from three elementary schools within a local inner-city district. These schools are located in a low to middle SES area with a high rate of transition and a mixed population. Approximately 90 percent of the students were of Hispanic background, 6% were Black, 2% were Asian, and 2% were non-minority Whites. There were 79 students classified as FEP (Fluent English Proficient) and 49 classified as LEP (Limited English Proficient). Language classification was indicated by the district, based on district reclassification criteria of language proficiency tests, achievement tests and teacher judgment.

Two forms of the diagnostic test were prepared. Both contained the same items but the order of the items was inverted: items that appeared on the first half of Form A were placed on the second half of Form B and vice versa.

Staff researchers were trained to administer the test. The test instructions allowed the administrators to clarify the meaning in vocabulary item stems but not in item distractors. The tests were administered at the schools. Students were allowed up to 90 minutes to complete the test although most students finished the test in about 45-60 minutes. Classroom teachers were presented during testing.

Overview of the Analytic Approach

The test score that a teacher uses to evaluate students' grasp of a curricular unit is typically the total score. If the whole class does poorly on a test of fractions, the teacher may decide to spend more time on the unit. If some students in the class do poorly on the test, the teacher may provide them with remedial instruction. Traditional approaches to reliability in educational and psychological measurement concern the dependability of that total score. The approaches focus on the consistency of students' scores over time (test-retest reliability) or from one test form to another (parallel forms reliability), or focus on the consistency of students' performance across items or sections of a test (internal consistency reliability).

Traditional Approach to Internal Consistency

Of the traditional approaches to reliability, only internal consistency reliability addresses the variability of performance across items within a test. Internal consistency alpha, for example, indicates how consistent student performance is across all items in a test. The magnitude of the coefficient shows whether the rank-ordering of student performance is stable across all items. A high value of alpha (at or near 1.00) indicates that the students who perform better than other students on one item also do so on the other items. A low value of alpha (at or near zero) indicates that the students who perform best on some items are not the same students who perfor best on other items. The latter result suggests that all items on the test are not measuring the same construct, and that student performance is different across different parts of the test. In this situation, the total test score is probably a poor indicator of students' mastery of the material.

While traditional approaches to internal consistency reliability provide some information about the consistency of performance across items in a test, they have limited usefulness for diagnosing specific areas of difficulty. For diagnostic purposes, it is important to have information about student performance on different parts of the test, i.e., a profile of scores. In the test of pronouns developed in the current study, it would be possible to obtain separate scores for each each rule of speech (nominative, objective, etc.), for singular and plural items, for first and third person items, and for each form of item (embedded in multiple sentences or non-embedded). While it would be possible to obtain such a detailed profile of scores for each student, this level of detail may not be necessary and might not be worth the cost of obtaining it. The central question is what level of detail in a profile is necessary to inform a teacher about difficulties that individual students or groups of students are having with the material.

The Analytic Questions

The analytic approach used in the present study focuses on the consistency of students' performance across multiple dimensions of a test, each dimension designed to measure a different aspect of the curricular unit. The aim of the analysis is to determine the minimum amount of information about student performance on the test that needs to be presented to guide teachers' future instructional decisions for individual students or for groups of students. The analysis addresses three issues: (1) the necessity of computing profiles of scores for individual students rather than only one for the class (or one for each subgroup of students in the class), (2) the level of detail that is necessary in the group or individual profiles, and (3) the number of items that are needed to obtain reliable scores in a profile.

Regarding the first issue, if all students have difficulty with the same material (for example, all students misunderstand how to use possessive pronouns), then a single profile for the whole class may be sufficient for diagnosing areas of difficulty. If some material is particularly troublesome to some students but is not troublesome to other students, then profiles for individual students may be necessary. Regarding the second isssue, if students perform equally well on all rules (nominative, objective, possessive), then it would not be necessary to provide separate scores for each rule. If, on the other hand, mastery of nominative pronouns is much greater than that of possessive pronouns, then it would be necessary to include in the profile separate scores for each rule. Regarding the third issue, once it is determined what scores should compose a profile, the question remains about the number of items that are needed to reliably measure each skill represented in the profile.

Multidimensional Approach to Test Structure: Generalizability Theory

Sketch of generalizability theory. To address the above issues, performance on the pronoun test was analyzed using generalizability theory. Generalizability (G) theory is a measurement theory designed to assess multiple sources of variation in a measurement (see Cronbach, Gleser, Nanda, & Rajarantnam, 1972; Shavelson & Webb, 1981; Webb & Shavelson, 1981). In a nutshell, G theory uses analysis of variance to partition sources of variation in measures of performance of behavior. The results of a generalizability study show the relative magnitudes of the sources of variation in a test and can be used to improve its design.

A measurement is a sample from a universe of admissible observations, characterized by one or more sources of error variation or facets (e.g., items, rules of grammar). This universe is typically defined as all combinations of the levels (called conditions in G theory) of the facets. Since different measurements may represent different universes, G theory speaks of universe scores rather than true scores, acknowledging that there are different universes to which decision makers may generalize. Likewise, the theory speaks of generalizability coefficients rather than the reliability coefficient, realizing that the value of the coefficient may change as definitions of the universe change.

In G theory, a measurement is decomposed into a component for the universe score and one or more error components. As an illustration, consider a 10-item test of pronoun knowledge in which 5 items measure singular pronouns and 5 items measure plural pronouns. This test has two facets: pronoun number (singular vs. plural) and item. If 20 students take this test, then the design underlying this study is a two-facet partially nested design with items (i) nested within pronoun number (n) and crossed with student (s). The object of measurement, here students, is not a source of error and, therefore, is not a facet.

The variance of the observed scores on this test (over all students and all items for each pronoun number) can be decompsed into independent sources of variation due to differences between students, items, and pronoun number and the interactions among them using analysis of variance. From the analysis of variance, an estimate of each component of variation in the scores is obtained: $\hat{\sigma}^2_s$, $\hat{\sigma}^2_n$, $\hat{\sigma}^2_{i,ni}$, $\hat{\sigma}^2_{sn}$, and $\hat{\sigma}^2_{si,sni,e}$. (Since items are nested within pronoun number in this design, the main effect for item (i) is confounded with the interaction between item and pronoun number (ni).) G theory focuses on these variance components. The relative magnitudes of the components provide information about particular sources of variation influencing performance on the test. The estimated variance component for students, $\hat{\sigma}^2_s$, is the universe score variance and is analogous to the true score variance in classical theory. The remaining variance components are considered error components.

G theory recognizes that decision makers (teachers, for example) may use the same score in different ways. Some interpretations focus on individual differences (relative decisions). For example, the teacher may be concerned mainly with the generalizability of the rank ordering of students, in order to give remedial instruction to the ten lowest-scoring students. Other interpretations may focus on the level of student performance itself, without reference to other students' peformance (absolute decisions). For example, the teacher may be concerned about a student's absolute level of pronoun knowledge, not how well he or she does relative to other students in the class.

Measurement error is defined differently for each of these proposed interpretations. For relative decisions, the error variance consists of all variance components representing interactions with the object of measurement (here, students):

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{sn}}{n_n} + \frac{\hat{\sigma}^2_{si,sni,e}}{n_i n_n}$$

In the above equation, n_n is the number of levels of the pronoun number facet and n_i is the number of items per pronoun number. The error variance for relative decisions reflects differences in rank ordering of students across items and pronoun number. If an interaction effect is large, then students' scores are not rank ordered the same across levels of the fact. For example, if the component representing the interaction between students and number is large relative to the other components, then students who perform the best on singular items are not the same students who perform the best on plural items.

For absolute decisions, the error variance consists of all variance components except that for universe scores:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{n}}{n_n} + \frac{\hat{\sigma}^2_{i,ni}}{n_i n_n} + \frac{\hat{\sigma}^2_{sn}}{n_n} + \frac{\hat{\sigma}^2_{si,sni,e}}{n_i n_n}$$

The error variance for absolute decisions reflects differences in mean performance of students across items and pronoun number as well as differences in rankings of students. When the decision maker is concerned with the absolute level of student performance, the variance components associated with effect of pronoun number and items ($\hat{\sigma}_n^2$ and $\hat{\sigma}_{i,ni}^2$) are included in error variance. The difficulty of one item as compared with another will influence a person's score. A test composed of easy items will suggest a higher level of proficiency than a test composed with difficult items. A large component for pronoun number, as another example, indicates that students find items one of number (say, plural) more difficult than items of the other number (singular).

Generalizability theory and score profiles. The relative magnitudes of the variance components contributing to relative error variance and absolute error variance can be used to determine what kinds of score profiles are necessary for diagnostic purposes. Wherever variance components contributing to relative error variance (interaction with students) are large, separate profiles are necessary for diagnosing learning difficulties. If the interaction between students and pronoun number is large, separate profiles would show which students were having more difficulty with plural items than singular items and which students were having more difficulty with singular items than with plural items. If the variance components contributing to relative error (interactions with students) are small, but the remaining components that contribute to absolute error (components that do not involve interations with students) are large, then one profile for the class would be sufficient. For example, if all students find plural items more difficult then singular items (a large variance component for pronoun number, $\hat{\sigma}^2_n$), than a profile for the class (the means for singular items and plural items) would show the average difference between plural and singular items. Finally, if the variance components that contribute to relative error variance and absolute error variance are both small, then student performance does not vary across the dimensions of the test. In this case, the total score on the test would be sufficient to guide decisions about instruction.

The above description concerns the relative magnitudes of the variance components, that is, the proportion of total variance accounted for by each variance component. A difficult decision is what proportion to be considered large. There is no rule of thumb about what proportion should be considered large. In the present study, all variance components that account for at least 3.5 % of the total variation will be noted and discussed. This level is conservative; other researchers might set a level of 5% or even 10% as the minimum proportion that should be used. As in all decision studies, there is a trade-off between cost and efficiency and information. Using a small proportion as a minimum may produce more detailed profiles than are necessary. Using a large proportion as a minimum, on the other hand, may cause important sources of variation to be overlooked or disregarded.

The optimal number of items in a profile. While stressing the importance of variance components and error variances, G theory also provides a coefficient analogous to the reliability coefficient in classical theory. The generalizability coefficient for relative decisions is defined as:

$$\hat{\rho}^2_{\text{Rel}} = \frac{\hat{\sigma}^2_s}{\hat{\sigma}^2_s + \hat{\sigma}^2_{\text{Abs}}}$$

An analoguous coefficient can be defined for absolute decisions:

$$\hat{\rho}^2_{\text{Abs}} = \frac{\hat{\sigma}^2_s}{\hat{\sigma}^2_s + \hat{\sigma}^2_{\text{Abs}}}$$

The generalizability coefficient, $\hat{\rho}^2$, indicates the proportion of observed score variance ($\hat{\sigma}^2_s + \hat{\sigma}^2_{\text{Rel}}$ or $\hat{\sigma}^2_s + \hat{\sigma}^2_{\text{Abs}}$) that is due to universe score variance ($\hat{\sigma}^2_s$). As the number of observations per student increases (for example, the number of items), the error variance ($\hat{\sigma}^2$ or $\hat{\sigma}^2_{\text{Abs}}$) decreases and the generalizability coeficient ($\hat{\rho}^2$) increases.

In the present context, the generalizability coefficient is useful for determining the number of items needed to provide a generalizable measure of each score in a profile. If the relative magnitudes of the variance components show that separate scores are needed for each student for plural pronouns and for singular pronouns (indicated by a large interaction between students and pronoun number), then one generalizability analysis would be performed for plural items and another one would be performed for singular items.

The design of each generalizability analysis is simple: student crossed with item. This design has three variance components: one for students (σ_s^2) , one for items (σ_i^2) , and one for the interaction between students and items plus unexplained residual variation $\sigma_{si,e}^2$. The error variance for relative decisions is:

$$\hat{\sigma}_{Rel}^2 = \frac{\hat{\sigma}_{si}^2}{n_i}$$

and the error variance for absolute decision is:

$$\hat{\sigma}_{Abs}^2 = \frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_{si}^2}{n_i}$$

If the analysis shows that a suitable level of generalizability (say, .70) can be obtained with 10 items, then the test would include 10 plural pronoun items and a student's mean on these 10 items would constitute his or her score for plural items in the profile.

If the variance components indicate that a profile of group mean scores is appropriate, then the object of measurement is the group, not the student, and the analysis changes accordingly (see Kane & Brennan, 1977). In the illustration used in the present study, there are two groups of students defined by their language backgroud: fluent English proficient and limited English proficient. In determining the mean score for a language group, the object of measurement is the language group. So the estimated variance component for language background $(\hat{\sigma}^2_1)$ is the universe score variance. The variation among students is error variation and so student becomes a facet of error. The design of the generalizability analysis of the number of items needed to measure a score in the group mean profile is students (s) nested within language group (1) and crossed with item (i). The error variance for relative decisions is:

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{s,sl}}{n_s} + \frac{\hat{\sigma}^2_{si,sli,e}}{n_s n_i}$$

and the error variance for absolute decisions is:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{i}}{n_i} + \frac{\hat{\sigma}^2_{li}}{n_i} + \frac{\hat{\sigma}^2_{s,sl}}{n_s} + \frac{\hat{\sigma}^2_{si,sli,e}}{n_s n_i}$$

If the analysis shows that 10 items are needed to produce a dependable measure of the group's knowledge of plural items, then the test should have 10 plural items.

Summary. In summary, the issues of the appropriate score profiles for diagnostic purposes and the number of items needed to produce dependable measures of each score in the profile are addressed in two stages. The first stage is a generalizability study of the structure of the test. In the illustration presented in this paper, the facets include: rule of grammar, pronoun number, context (embedded vs. non-embedded), person (first person vs. third person, form (relative vs. non-relative), and item. The relative magnitudes of the variance components in this design show which score should be included in individual student profiles. The second stage is a separate generalizability analysis for each skill in the individual and group profiles to determine the number of items that should be included in the test to obtain a dependable measure of those skills.

Results of Illustrative Analyses

This section illustrates the analytic approach to diagnostic testing described in the previous section. It summarizes (1) the preliminary analyses to determine which population subgroups to include in the generalizability analyses; (2) the three designs underlying the generalizability analyses of test structure; (3) the variance components produced by the generalizability analyses, (4) example diagnostic profiles; and (5) the number of items that would be needed to yield dependable measures of each score in the diagnostic profiles.

Preliminary Analyses

The first step in the approach to diagnostic testing presented here is to determine whether there are distinct population subgroups in the design. In the present illustration, the pronoun test was administered to students from multiple classrooms and schools, and students differed in ethnic background, language background, and age. Therefore, preliminary analyses were conducted to determine whether these factors influenced performance on the pronoun test. Analysis of variance F tests revealed that the only population characteristic influencing performance on the test was language background (FEP vs. LEP; $F(1) = 30.09$, $p < .001$). The statistical tests for classroom, school, ethnic background, and age were not significant (F statistics ranged from .12, $p < .73$ to 1.06, $p < .37$. In all further analyses, then, only the distinction between FEP and LEP students was maintained.

Summary of the Three Designs

As was described in the section summarizing the design of the test, the entire test can be described by three crossed designs. Design I is a five-facet design yielding 64 items: embeddedness (2 levels), pronoun form (2 levels), rules (4 levels), number (2 levels), and 2 items for each combination of the previous four facets. Design II is a four-facet design yielding 40 items: embeddedness (2 levels), rules (5 levels), number (2 levels), and 2 items for each combination of the previous facets. Design II is also a four-facet design yielding 40 items: embeddedness (2 levels), rules (5 levels), person (2 levels), and 2 items for each combination.

As a result of the complexity of each design, the number of variance components in each analysis was very large. For example, in the analysis of Design I, with students nested within language background and students and language background crossed with embeddedness, rules of grammar, number, form, and item, there were 51 variance components. Rather than present the descriptive results (means, standard deviations) for all variance components in each design, descriptive results are presented only for components that account for at least 3.5% of the total variation in the design. Table 2 presents the variance components that exceed 3.5% of the total variation in each of the three designs. Each number in Table 2 represents the percentage of total variation accounted for by each variance component. Tables 3, 4, and 5 present the means and standard deviations corresponding to all variance components listed in Table 2. The means are the percent correct, so the maximum score possible is 1.00.

Variation due to student and language background. The large component for language background in each design indicated that FEP and LEP students showed different levels of performance on the test. As the descriptive results in Tables 3 through 5 show, FEP students showed higher mastery of pronoun usage than did LEP students. The large variance component for students (nested within language background) in all three designs shows that there were substantial individual differences between students within a language group. Some students had mastered pronoun usage while others had not. The component for students, then, reflects the range of mastery of pronoun usage in the same.

Variation contributing to absolute error. Most of the variance components presented in Table 2 do not involve interactions with students or with language background. In Design I, the pronoun form facet accounts for the greatest variance (34.0%). Students found relative pronoun items to be very difficult. In fact, as can be seen in Table 3, LEP students performed at about chance level on all relative pronoun items except those measuring the nominative rule (with 5 response choices for each item, chance level is 20%).

Table 2 also shows a substantial effect for the context of the item, whether the sentence was embedded with a paragraph. The

Table 2
Proportion of Total Variation accounted For by Each Variance Component

Variance Component	Design I	Design II	Design III
Language Background [L]	12.3	19.6	19.4
Student [S(L)]	20.6	33.6	37.3
Pronoun Form [F]	34.0	—[a]	—
Embedded [E]	4.1	26.2	17.3
Rule [R]	<3.5	<3.5	6.0
F E	5.1	—	—
F R	4.6	—	—
F S(L)	4.0	—	—
E S(L)	<3.5	6.6	<3.5
Residual	3.7	5.6	6.1
All others	11.6	8.4	13.9
Total	100.0	100.0	100.0

[a] Not applicable.

Note: Only variance components accounting for more than 3.5% of total variance are listed here.

variance component for embeddedness is smaller for Design I than for Designs II and III because the effect of relative pronouns (who-whom) overwhelmed that of embedding in Design I. The means in Table 3, 4, and 5 show that all students found it much more difficult to determine correct pronoun usage when the target sentence was embedded within other sentences. The difference in performance between embedded and non-embedded items was similar for FEP and LEP students.

Interestingly, the rule of grammar produced substantial variation in performance only in Design III. As Table 5 shows, students tended to perform worse on the items measuring the possessive rule than on items measuring the other rules. This effect appeared only when items measuring plural pronouns were included in the analysis (Design III), and not when singular items were included (Design II). As is indicated by the small variance component for rule in Design I (where the possessive rule was not included), student performance did not vary much across item measuring knowledge of the nominative and three objective rules.

Table 2 shows two other effects in Design I that contributed to absolute error variance. The pronoun form facet interacted with the embeddness facet and with rules. The interaction between pronoun form and embeddedness indicates that the difference between performance on embedded and non-embedded items was not constant across relative and non-relative pronoun items. This result is clearly seen in Table 3. Both FEP and LEP students did much better on non-embedded items than on embedded items only when the pronouns were not in the relative

Table 3
Descriptive Results for Major Sources of Variation
in Design I

Factor	FEP[a]		LEP[a]	
	M	SD	M	SD
Non Relative Pronouns	.70	.15	.54	.18
Context				
Non-Embedded	.87	.17	.66	.24
Embedded	.53	.22	.42	.20
Rule				
Nominative	.59	.21	.49	.19
Direct Object	.66	.23	.51	.21
Indirect Object	.76	.23	.56	.28
Object of Preposition	.78	.18	.60	.29
Relative Pronouns	.34	.15	.24	.13
Context				
Non-embedded	.35	.17	.27	.14
Embedded	.33	.18	.22	.15
Rule				
Nominative	.57	.24	.52	.24
Direct Object	.20	.18	.13	.17
Indirect Object	.25	.21	.19	.21
Object of Preposition	.33	.28	.14	.20

[a] FEP = Fluent English Speaking. LEP = Limited English Speaking.

Table 4
Descriptive Results for Major Sources of Variation in Design II

Factor	FEP		LEP	
	M	SD	M	SD
Context Non Embedded	.84	.17	.62	.24
Embedded	.51	.21	.38	.18

Table 5
Descriptive Results for Major Sources of Variation in Design III

Factor	FEP		LEP	
	M	SD	M	SD
Context				
Non-Embedded	.78	.17	.61	.23
Embedded	.56	.20	.42	.19
Rule				
Nominative	.62	.20	.49	.23
Direct Object	.68	.22	.53	.21
Indirect Object	.77	.23	.64	.25
Object of Preposition	.81	.20	.63	.31
Possessive	.46	.25	.27	.21

form. When the items called for relative pronouns, in contrast, student performance was very similar for non-embedded and embedded items. Thus, for relative pronoun items, the presence or absence of context did not affect student performance. Students performed poorly in both cases.

The interaction between pronoun form and rules can also be seen clearly in Table 3. For non-relative pronouns, students showed similar performance on all rules, with performance on nominative items somewhat lower than performance on objective items. For relative pronouns, on the other hand, performance on the nominative items was much higher than performance on the objective items. About half of the students knew when to use "who" (the relative form of nominative pronouns) but very few FEP students and no LEP students knew when to use "whom" (the relative form of objective pronouns). To determine whether students performance differed across the three objective rules, Design I was also analyzed without the nominative rules (including only the three objective rules). The interaction between pronoun form and objective rules nearly disappeared (it accounted for only about 1% of the total variance), showing that student performed nearly the same on the three objective rules. Given this finding, then, it would not be necessary to retain information on the three objective rules. The mean for all objective items as an undifferentiated set would be sufficient.

Variation contributing to relative error. A notable feature of Table 2 is the lack of interactions between any facet and language background. This finding shows that the pattern of performance across the dimensions of the test among FEP students was the same as that for

LEP students. Coupled with the large component for language background, this result indicates that the profiles for the two groups have the same shape, with the profile for FEP students being higher than that for LEP students.

There were surprisingly few interactions between students and facets. The component for the interaction between students and pronoun form in Table 2 indicates that the rank order of students on relative items was not the same as the rank order of students on the non-relative items. There are two possible interpretations of this result. The first, which is highly unlikely given the huge main effect for pronoun form, is that some students found the relative pronoun item easier than the non-relative pronoun items while the rest found the non-relative pronoun items easier than the relative pronoun items. A far more likely interpretation is that the difference in performance between relative and non-relative pronouns was larger for some students than for others. It is unlikely that any students performed better on relative pronouns than on non-relative pronouns.

A similar interpretation can be given for the interaction between students and the embeddedness facet in Design II. Since it is unlikely that any student performed better on embedded items than on non-embedded items, the most likely interpretation of the interaction is that the difference in performance between embedded and non-embedded items was larger for some students than for others.

Finally, it should be noted that the residual variance component represents the interaction between all facets in the design, including students and language background, plus unsystematic error. A large residual variance component usually reflects sources of variation that have not been taken into account in the measurement. The small magnitude of the residual component in all three designs in the present study suggests that all important test facets have been taken into account in the design of the test.

The Primary Sources of Variation and Example Diagnostic Profiles

The only sources of variation in test performance that exceeded 3.5% of the total variation were the pronoun form, embeddedness, and rule facets. The person (first vs. third) and the number (singular vs. plural) of the pronoun did not produce variation among students' test scores. That is, students showed equal mastery of first and third person pronouns and showed equal mastery of singular and plural pronouns. Furthermore, the effect for items was very small, indicating that students performed similarly on both items in each cell of the test design.

The findings protrayed in Table 2 and described above can be used to make recommendations about the optimal diagnostic profiles for pronoun usage for the sample in this illustrative study. Only the large effect contributing to relative error (those involving interactions with students) would need to be incorporated into the score profiles for individual students. Since only the who-whom and embeddedness facets interacted with students, the profile for individual students would need only to consist of the mean scores for relative pronoun items, non-relative pronoun items, embedded items, and non-embedded items. Example profiles for three randomly selected students appear in Figure 1.

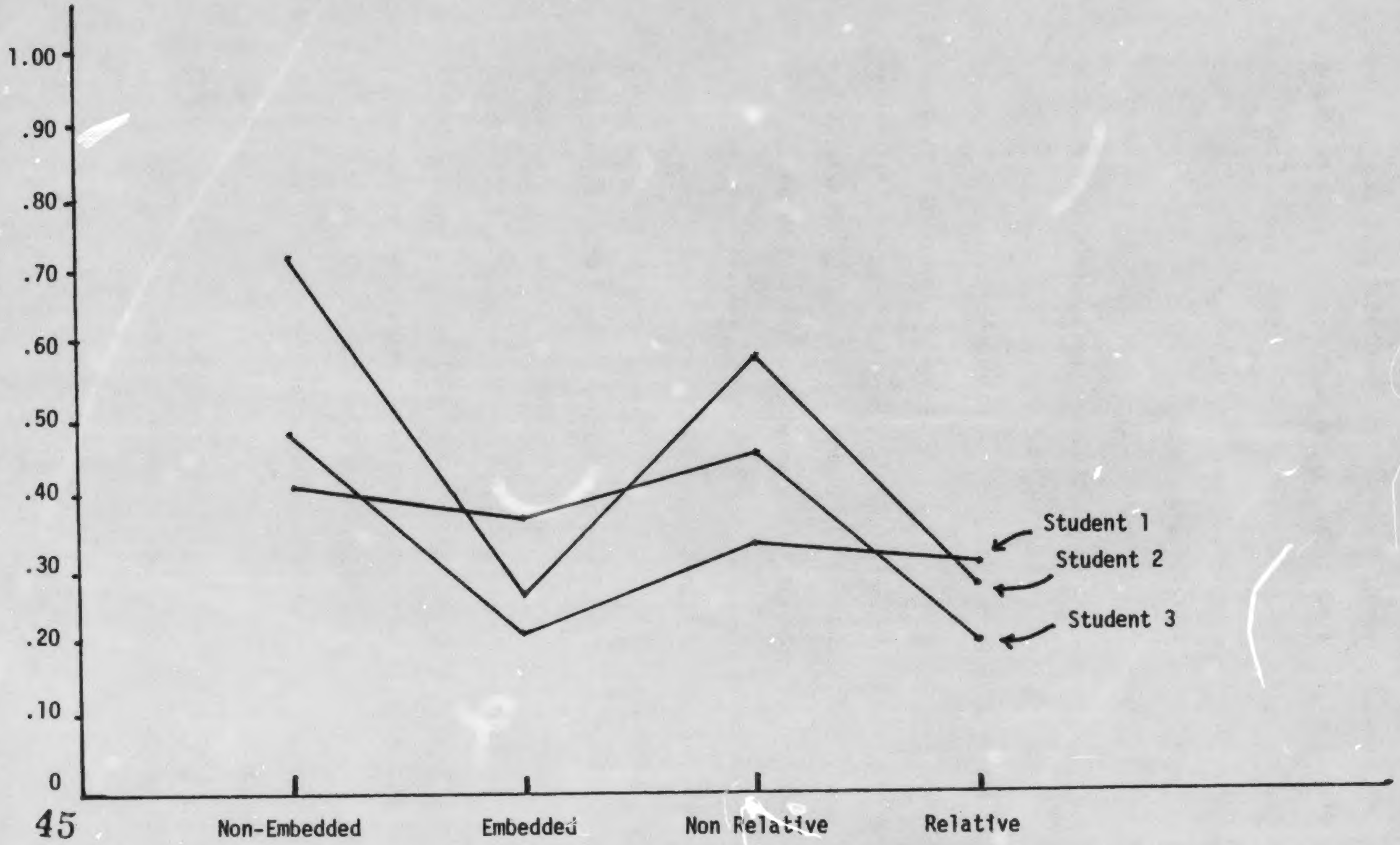

Figure 1. Individual Profiles for Three Students

The remaining large variance components--those contributing to absolute error but not relative error (components that do not involve interactions with students)--would guide the formation of class or group profiles. Since pronoun form interacted with embeddedness and rules, the group profile should present the means for embeddedness (embedded, non-embedded) and rules (nominative, objective, possessive) separately for relative items and non-relative items, as was done in Table 3. Figure 2 presents such profiles for FEP and LEP students. Since performance was similar across the three objective rules, only the mean socre is presented for objective items. Furthermore, since performance was similar across person and number, first person plural, third person singular, and third person plural items were combined.

The mean profiles in Figure 2 show the general patterns of performance in this sample. Since the rule facet in the design did not interact with student, the means for nominative, objective, and possessive items are good representations of the performance of all students. This profile would show that all students need further instruction on the possessive rule and the objective rule for relative form. Similarly, the general pattern for embeddedness in Figure 2 suggests that students would need further instruction on all embedded pronouns and non-embedded relative pronouns.

In summary, the variance component analyses in the present study show that individual profiles for embedded and non-embedded and relative and non-relative pronouns, and a group profile for rules of grammar would be sufficient for diagnosing individual and group difficulties with pronoun usage. These profiles would be more

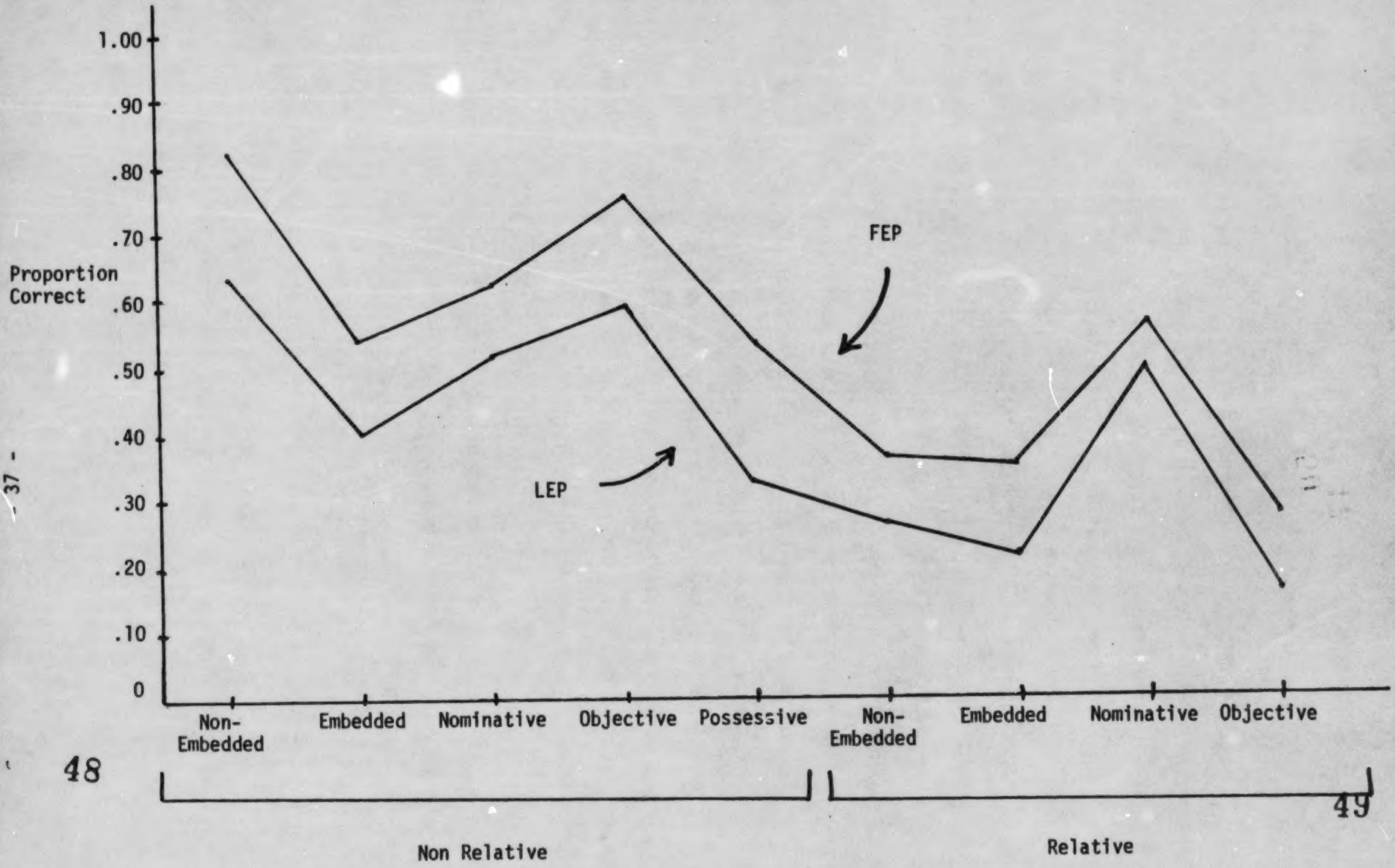

Figure 2. Mean Profiles for FEP and LEP Students

informative than the total score of the test, and suggest that diagnostic decisions based only on the total score might lead to erroneous consequences for the student and for the class. Not only do the variance component analyses show which aspects of test performance should be tabulated for individual and group diagnosis, they are also valuable for showing which aspects of pronoun usage do not need to be tabulated. Since student performance did not vary across number and person, these facets could be omitted from the diagnostic profiles.

The Optimal Number of Items

The previous section demonstrated how to use the relative magnitudes of the variance components to guide slection of scores for student and group profiles. This section reports the results of generalizability analyses to show how many items would have to be included in the test for dependable profile scores. The design of the generalizability analysis for each of the four scores in the individual profile (non relative, relative, non-embedded, embedded, see Figure 2) was students crossed with items. The items used in each generalizability analysis were all items in the original test that pertained to that score. The generalizability analysis of non-embedded items, for example, included all first-person, third-person, singular, plural, relative, non-relative, nominative, objective, and possessive items that were non-embedded.

The results of the generalizability analyses are presented in Table 6. Table 6 shows the number of items corresponding to different levels of generalizability. For example, it would take at least 10

TABLE 6

Number of Items Corresponding to Different Generalizability Coefficients

	Relative Decisions				Absolute Decisions			
Score	.50	.60	.70	.80	.50	.60	.70	.80
INDIVIDUAL PROFILE								
Non-Embedded	8	10+	10+	10+	10+	10+	10+	10+
Embedded	9	10+	10+	10+	10	10+	10+	10+
Non Relative	5	9	10+	10+	7	10	10+	10+
Relative	10	10+	10+	10+	10+	10+	10+	10+
GROUP PROFILE								
Non Relative								
Non-Embedded	1	1	1	1	1	1	2	3
Embedded	1	2	3	7	6	9	10+	10+
Nominative	2	3	4	9	10+	10+	10+	10+
Objective	1	1	1	2	3	4	6	10
Possessive	1	1	1	1	2	2	4	6
Relative								
Non-Embedded	2	3	4	9	10+	10+	10+	10+
Embedded	1	2	2	4	4	7	10	10+
Nominative	1	1	1	1	2	2	3	5
Objective	1	2	2	5	2	2	4	8

[a] More than 10 items would be needed to obtain this level of generalizability.

items to measure individual proficiency on non-embedded pronouns with a .70 level of generalizability for relative and absolute decisions. On the other hand, only one item would be needed to measure group mean proficiency on non-embedded pronouns at the same level of generalizability for relative decisions (two items would be needed for absolute decisions).

Since the same items can be used to measure different aspects of pronoun usage, the total number of items in the test needed to obtain generalizable measures of each score in the profile is smaller than the sum of the number of items in Table 6. For example, an embedded, relative nominative item can be used to measure embedded pronoun usage, relative pronoun (who, whom) usage, and nominative pronoun usage. If the scores in the individual and group profiles were to be used for relative decisions (for example, selecting the bottom 20% of students for remedial instruction), a pronoun test of 20 items could be constructed so that each score in the profile had at least .70 generalizability. A pronoun test with the following configuration would satisfy the requirement listed in Table 6: 4 non who-whom objective (embedded) item, 1 non who-whom possessive (embedded) item, 1 who-whom nominative (non-embedded) item, 2 who-whom objective (embedded) items, 3 other who-whom non-embedded items, 3 other embedded items (all who-whom) and 5 other non-embedded items (1 whom-whom and 4 non-who-whom). For absolute decisions, a pronoun test with 40 items could be constructed so that each score in the individual and group profiles had a level of generalizability of .70.

Distractor Analysis

Test subscores based on simple right-wrong scoring, such as those reported above, utilize only a portion of the data that may be available from student responses. The nature of their incorrect response provides potentially useful additional information for diagnosing student needs. In the case of multiple choice items where distractors are constructed to represent specific errors or misconceptions, analysis of distractor choices might supply several pieces of information:

First, such an analyses can point to potential reasons for students' difficulties with certain subscales by identifying the kinds of errors they made. For example, the distractor analysis of the relative pronoun items can show:

> Whether student responses are systematic or random: e.g., did they always choose between who and which, or was there another popular response? Were they consistently choosing among these distractors?
>
> Whether the kinds of errors students make are due to a misconception of a rule or due to some other feature. For example, if students consistently choose who when whom is the correct response, they do not know the rule which regulates the use of these two pronouns. If there is another frequently selected distractor such as "he", they may be exhibiting an additional problem in the construction of independent clauses.

Secondly, distractor analysis can indicate error patterns across subscales. These patterns can point to erroneous rules, misconceptions or misinformation which are applicable to several or all of the subscales. For example, one of the error patterns that occurred for the relative pronoun items might also appear in items testing other types of pronouns and suggest that students are having difficulty with a particular aspect of syntax.

Method

Recall that item distractors are organized such that three distractors represent a specified error and one distractor reflects guessing. Matched pair of item contain the same distractors. For example:

Cornelia is a very pretty woman.	David is a handsome man.
him (incorrect gender & pronoun)	her
they (incorrect case)	they
she (incorrect response)	he
he (incorrect gender)	she
her (incorrect pronoun type)	him

Each item was analyzed to determine what percentage of test takers selected each distractor; and contingency analyses were conducted to examine consistency of student responses. Responses were to be analyzed to address a number of enter-related questions:

1. Were the response patterns within matched pairs of items consistent? That is, did students select the same distractors for both items?
2. Did students make the same error across several types of items or were some errors characteristic of one type of item?
3. We know that the pattern of responses differed for non-embedded and embedded items; can distractor analysis point to potential reasons for this pattern?
4. Did Limited and Fluent English Proficient students (LEP & FEP) differ in their response patterns? Did patterns indicate that they were making different or similar types of errors.

Results

The validity of an analysis of distractor patterns rest on some consistency in student wrong answer choices. When distractor patterns are consistent within a matched pair it seems reasonable to assume that the error patterns reflect the misconceptions or misinformation specified by the distractors. In the absence of such consistency, then it is unclear whether students are selecting at random, are reacting to some pecularity in item content or to linguistic or other properties.

Contrary to expectation, contingency analyses found little consistency in student wrong answer choices. Distractor selection was consistent, on the average, in only 21% of the cases, ranging from 0% to 48% over the 42 parallel pairs of items included on the test; two-thirds of the items fell in the 11-30% range (See Table 7). There was greater consistency among more difficult items, but no apparent differences between fully and limited English proficient students. No patterns were discernible by pronoun rule.

Table 7
Agreement on distractor choices
for parallel pairs of items
Wrong Answers Only

	Limited English Proficient		Fluent English Proficient	
Mean % of agreement	frequency	%	frequency	%
0-10%	7	(15)	7	(15)
11-20%	13	(28)	16	(35)
21-30%	19	(41)	15	(33)
31-40%	5	(12)	6	(13)
40-50%	2	(4)	2	(4)
	46 pairs	(100)	46 pairs	(100)
	n=49		n=79	

Future analyses will examine alternative models for describing student error patterns, looking for instance at whether students are consistently eliminating certain distractors and guessing at random from those remaining. The results of these analyses will be included in a subsequent report.

Conclusions

This paper described a four-step approach to constructing a diagnostic test that provides precise but practical information on students' problems and needs for additional instruction or remediation. The approach is based on analyzing the structure of the domain to determine which skills within the domain need to be assessed to diagnose students' problems.

The first step in the diagnostic process described here was to identify the factors that described the curricular domain (here, pronoun usage). Four content factors were identified: the rule of grammar (nominative, objective, possessive), the pronoun form (relative--who or whom, non-relative), the number (singular, plural), and the person (first, third). In addition, a factor corresponding to cognitive complexity was identifed: whether the context of the reading passage had to be taken into account to determine the correct pronoun. This factor was operationalized in two levels of embedding: a single sentence or a pragraph.

The second step was to construct a test with items representing all possible combinations of factors (content and cognitive complexity). Sensible items could be written for 46 combinations of factors. Two items were written for each combination, resulting in a 92 item test.

The third step in the diagnostic testing process used generalizability theory to determine which factors and interactions among them produced variation in students' scores. Specifically, the relative magnitudes of the variance components corresponding to all factors and interaction among them in the test revealed which factors were important. This information was used to identify the information needed in diagnostic profiles. Only two content factors, rule and pronoun form, produced variation in student performance. The other two content factors, number and person, did not. Furthermore, cognitive complexity also had a large effect on student performance.

Some difficulties were common to all students (e.g., all students had more difficulty with possessive pronouns than with objective pronouns). This information could be entered in a single profile for the group or class. Other difficulties applied to some students but not others (e.g., some students did much worse on embedded items than on non-embedded items while other students performed similarily on both types of items). This information would be part of profiles for individual students. Since the number and person factors had no effect on student performance--all students performed about the same on singular and plural items on first-person and third-person items--there was no need to distinguish between these skills in the test or in the profiles.

Based on the information about the necessary ingredients of diagnostic profiles, the final step in the analytic process was to determine the minimum number of items needed to obtain generalizable measure of each skill in the diagnostic profile. The results of the generalizability analyses showed that a 20-item test would be sufficient to measure mastery of pronoun usage if the teacher's interest was in identifying the students with greatest need for additional instruction in each skill. A 40-item test would be sufficient if the teacher's interest was in identifying each student's absolute level of mastery of each skill.

In short, the structure of the domain consisted of 46 skills in pronoun usage (all sensible combinations of the five factors). The inital test consisted of 92 items, 2 per skill. To adequately measure student performance on each of these 46 skills would probably take between 2 and 10 items per skill, resulting in an extremely long test. The analyses performed here showed that only 9 of the 46 skills need be assessed resulting in a vastly simplified and shorter diagnostic test.

Although the entire process of (1) identifying a domain, (2) constructing an inital test to fully represent the domain, (3) analyzing the performance on the inital test to determine the factors that influence student performance, and (4) constructing the final optimal test would be too time-consuming for a classroom teacher, the use of the final diagnostic test and score profiles would certainly be feasible for classroom practice. With a relatively short test (maximum of 20 minute test to administer, in this case), the teacher could identify students' strengths and weaknesses on all important aspects of the curriculum domain and make instructional decisions accordingly.

Specification of the domain structure underlying the test in an important issue in this diagnostic approach. It is important to specify the test as completely as possible. If factors in the test are left out, difficulties that students have on the test may be attributed to the wrong skills or may not be able to be identified at all. Although complete specification is important, it is not necessarily difficult. In the present study, the generalizability analyses showed that only a small amount of variation in test performance was attributed to unexplained factors. Consequently, it is reasonable to conclude that all important factors in the domain were included.

Also important in domain specification is not to restruct the domain only to aspects of content. Although several content factors did affect student performance, the cognitve complexity of the item had a major impact on performance. For example, even though many students could correctly identify when nominative pronouns should be used in a single sentence (a low level of cognitive complexity), many of them could no do so when the sentence was embedded in a paragraph requiring them to use the context of the paragraph (a high level of cognitive complexity). A teacher would come to different conclusions about mastery of pronoun usage from a test with items of low cognitive complexity and from a test with items of high cognitive complexity. Without taking into account the influence of cognitive complexity on performance, the teacher may well make erroneous decisions about the need for additional instruction.

Finally, the results of the illustrative analyses presented here also have implications for taking into account multiple student populations. Teachers often give different tests to students from different population subgroups (for example, different language backgrounds), assuming that the performance of the groups is different. An implicit assumption, therefore, may be that some groups excel on some material while other groups excel on other material; that is, that profiles of different groups may have different shapes. The strikingly parallel profiles of fluent English proficient students and limited English proficient students in the present illustrative study, however, raises a question about whether different tests are necessary. In this case, separate tests for each group would be unnecessary. To take into account the mean differences in performance between groups (fluent and limited English proficient students), the items measuring a particular skill on the diagnostic test could cover a range of difficulty (for example, varying the vocabulary level, or length of the sentences in a item).

REFERENCES

Baker, E.L. Beyond objectives: Domain referenced tests for evaluation and instructional improvement. Education Technology, 1974, 14.

Baker, E.L., & Herman, J. Task structure design: Beyond linkage. Journal of Educational Measurement, Summer 1983, 20(2), 149-164.

Block, J.H. (Ed.). Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971.

Bloom, B.S. Human characteristics and school learning. New York: McGraw Hill Book Company, 1976.

Chi, T.H., & Glaser, R. The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E.L. Baker and E.S. Quellmalz (Eds.), Educational testing and evaluation. Design, analysis, and policy. Beverly Hills, California: Sage Publications, 1980.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Doehring, D.G. & Aulls, M.W. The Interactive Nature of Reading Acquisition. Journal of Reading Behavior, 79 vol 11, #9, pp 27-40.

Edmonds, R. Making Public Schools Effective. Social Policy, 1981, 12, 56-60.

Glaser, R. Evaluation of instruction and changing educational models. In M.C. Wittrock and D.E. Wiley (Eds.), The evaluation of instruction. Chicago: Rand McNally and Company, 1970.

Hambleton, R. Item selection methods with criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Hively, W., Maxwell, G., Rabehl, G. Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph No. I. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hunter, M. Mastery teaching: increasing instructional effectiveness in secondary school, colleges and universities. TIP Publications, El Segundo, California: 1983.

Kane, M.T., & Brennan, R.L. (1977). The generalizability of class means. Review of Educational Research, 47, 167-292.

Klausmeier, H.J. (Ed.) Individually guided education: 1966-1980. Journal of Teacher Education, 1976, 3, 199-206.

Millman, J. Computer-based item generation. In R.A. Berk (Ed.), Criterion-referenced measurement. Baltimore, Maryland: John Hopkins University Press, 1980.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychology Measurement, 1968, 28, 95-104.

Popham, W.J. Domain specification strategies. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: John Hopkins University Press, 1980.

Popham, W.J. & Baker, E.L. Classroom instructional tactics. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. 1976.

Quellmalz, E. Cognitive models for linking testing and evaluation. Los Angeles, California: Center for the Study of Evaluation, 1982.

Shavelson, R.J. & Webb, N.M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166, 1981.

Spiro, R.J. Constructive Processes in Prose Comprehension and Recall in S.R.J. Spiro, B.C. Bruce & W.F. Brewer (eds.), Theoretical Issues in Reading Comprehension Perpectives from Cognitive Psychology, Linguistics, Artificial Intelligence and Education. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Tatsuoka, K.K., Birenbaum, M., Tatsuoka, M.M., & Baillie, R. A psychometric approach to error analysis on response patterns. (Research Report 80-3). Urbana, Illinois: University of Illinois, Computer-based Education Research Laboratory, February 1980.

Webb, N.M., Shavelson, R.J. (1981). Generalizability of general educational development ratings of jobs in the U.S. Journal of Applied Psychology, 66, 186-191.

DIAGNOSING STUDENT ERRORS:
AN EXAMPLE FROM SCIENCE

by

Steve Shaha

Table of Contents

	Page
Diagnosing student errors: an example from science	1
1. Test Development	4
2. Item Design	6
Experiment I	9
1. Method	9
2. Results and Discussion	11
Experiment II	11
1. Method	11
2. Results and Discussion	12
Conclusions	13
References	18

Diagnosing Student Errors:
An Example From Science

Steve Shaha

Center for the Study of Evaluation

Most tests of science are designed to assess knowledge of scientific facts. Information gained from such tests generally is summarized in a score which reflects the number of correct responses made. Hence, any diagnostic information is relatively limited; the number of correct responses relates to knowledge of scientific facts alone and not to an interpretable summary of the nature of a student's misunderstanding. More often than not, total scores which are useful for assessing class standing are not very useful for helping a teacher to aid individual students in overcoming specific problems. Perhaps the most promising feature of diagnostic tests is that they will yield information above the level of mere number of correct versus incorrect responses. Well designed diagnostic tests would give specific information to instructors concerning the individual student's problems.

As part of a larger project in diagnostic assessment, the Center for the Study of Evaluation undertook the development of a prototype diagnostic test of understanding of science at the high school level. The test's primary purpose was to provide detailed information concerning strengths and weaknesses in student's skills in scientific reasoning. Our interest was a) to isolate specific errors of scientific reasoning which by their consistency could be systematically diagnosed, and b) to examine differential patterns of error at different levels of comprehension. The first factor necessitates a rational choice of distractors; the second necessitates an orderly structuring of items by complexity of item content.

Previous work in diagnostic testing in science is notable for its variety of almost completely independent derivations, in a number of countries around the world, of both theories of science understanding and tests which measure it. After the well-known but two-decade-old "Test on Understanding Science" (Cooley & Klopfer, 1961) and the massive "Test of Understanding the Nature of Science" (IEA, 1969), there appears to have been no new well-documented instrument in this area. However, independent workers have succeeded in certain areas with demonstration projects. Dillashaw and Okey (1980), from the US, developed a test for diagnosing five process skills in science. Their findings justified the use of diagnostic tests, showing that both achievement scores and attitudes toward science rose for high school students when information from diagnostic tests was used for remedial purposes. Gorodetsky and Hoz (1980), from Israel, explored the use of profile analysis techniques for diagnosing conceptual misjudgments in scientific logic. Information from this type of diagnostic test showed that certain conceptual problems associated with lags in science learning can be identified, and that remediation of these problems can lead to resolution of learning lags. Dreyfus and Jungwirth (1980a; 1980b), also from Israel, succeeded in isolating and classifying the actual types of errors most often committed in responding to questions concerning science. Their efforts showed that is was indeed possible to develop diagnostic tests of scientific reasoning based upon the assessment of patterns in errors committed.

Billeh and Malik (1977), from Jordan and Pakistan respectively, report success at the college level with an 85 item instrument which taps assumptions of science, scientific ethics, and other areas; the results might be applicable to a diagnostic interpretation but only summary scores are presented. Rubba & Anderson (1978), from the US, developed an instrument based on a nine-factor conceptualization of science undertanding by Showalter (1974); again diagnostic interpretations may be feasible but summary scores alone are discussed. Cantu & Herron (1978), from Mexico and the US respectively, present a science concept attainment test which allows an interpretation following a piagetian understanding of how students learn in that field. Osborne (1978), from New Zealand, explores subscales of a test of college-level physics which allow a diagnostic interpretation in each of seven curricular components. Rodrigues (1980), from Brazil, administers a set of cartoons of certain physical situations which reflect laws at work, and forms a diagnosis of sorts based on characterizing the narratives children give in response.

The best single theoretical contribution to diagnostic testing is by Johnstone (1981), from Scotland. He presents some selected examples of college level science tests which allow an interpretation of specific types of misunderstanding. He also discusses a variety of distorting factors which are germaine not only to science tests but to all educational tests from which a diagnosis might be derived. These distortions arise in the context of superficial testing (and so a test "must probe into each linkage and subconcept as well as take a global picture" p. 39), too much guessing (and so a test should allow confidence marking) and issues of language clarity. He takes the view that diagnostic tests in science

> are relatively easy to construct because course objectives can be specified for a major portion of the study. Items can be written to test each objective from several angles and the items can often be in an objective format which lends itself to computer marking. (p. 34).

The reported ease of construction would be a significant advantage if true but it remains to be studied in more detail.

Test Development

Because none of the extant tests of science understanding incorporate both a rational basis for distractor choice and a uniform layering of items by complexity, four topics areas within science were selected for a prototype diagnostic test of science: photosynthesis, magnetism, energy in a closed system, and gravity. The target audience was the high school level. It is important to note that since we could not control for instruction received by students tested, we selected areas of science which we considered to be generally covered from seventh grade upward, based on a search of commonly used textbooks.

Item generation for the present study was based on two parallel considerations, the definition of content and principles to be tested and the determination of the specific types of errors which the test would be designed to diagnose. On the matter of content, we decided to develop items which mirrored the level at which a given content area was understood -- that is, items which would provide a test of a student's depth of comprehension. To accomplish this we chose a three level comprehension strategy:

- * Low level -- Definition. Items designed to test for knowledge of a principle at the definition or factual level.
- * Middle level -- Principle Application. Items which required that subjects recognize the use of a principle in a given situation or context.
- * High level -- Problem Solving. items which required subjects to analyse a problem or probably unfamiliar situation and arrive at the correct solution by applying the principle in question.

The next step was to define the errors to be diagnosed. Diagnostic tests do not merely measure correct versus incorrect responding, but are specifically designed to assess well-defined errors in a systematic manner. The error types we selected for diagnosing in the four areas chosen were based on the research of Dreyfus and Jungwirth (1980). In their study, they were successful in isolating five classifications of error types which are common in scientific reasoning, of which the following errors were of interest in the present study:

Logical fallacy -- Student's response based upon faulty logic, such as circular reasoning, generalizing beyond the specific situation, or imputing cause from correlation.

Intuition -- Students response based upon intuitive logic or prior experience rather than a understanding of the problem.

Content -- Student's response selected due to similarity with the language in the stem. Also test wise errors fit in this category, such as selecting the most scientific-sounding alternative or the longest choice available.

The items generated for each topic area selected were based on crossing the two dimensions: each level of comprehension to be tested, with each error type to be measured. There were three items generated at each level of comprehension, and every error type was assessed three times at each given comprehension level. Strict procedures were followed in the establishment of domain specifications, rules for generating items, and the outlining of objectives to be tested (which are elaborated in the accompanying manual by McArthur, Shaha, Choppin, & Hafner [1983]).

Item Design

The item design process is summarized below. The most effective way to envision the process is to consider actual test items. For this purpose, three items from the area of Magnetism will be explored, beginning with the lowest level of comprehension.

The first item of interest is the item at the low level of comprehension, the item which should be easiest for the most subjects. This item was constructed to test the most basic principles of the concept, even at the level of rote learning. In magnetism, the following item stem was developed:

> Nails made of metal alloys (mixtures of metals) are attracted by magnets.

The stem requires only that a student verify the truthfulness or falsehood of the assertion presented. In this instance the distractors are a tool for sorting out the precise reasons for both correct or erroneous responses. Each distractor contained a possible rationale for deciding whether the statement was true or false, and these rationales were designed to provide information concerning the precise type of errors being committed. Of the four distractors per item, two allowed selecting True as a response, and two allowed False. The following were the four alternatives for the item above:

a. True -- because all metals are magnetic. [Logical fallacy]

b. True -- only if the alloy contains iron. [Correct response]

c. False -- because only iron can be magnetized. [Content distraction]

d. False -- because nails are not the same shape as magnets. [Intuition distraction]

The first alternative represents a logical fallacy, since selecting this alternative indicates that a generalizing from a single example (iron) to an entire universe of related examples (metals). The second alternative is correct. The third alternative is indicative of an error based on content because its wording is very similar to the correct response as well as the stem. The last alternative is an example of an intuition distraction, since its selection is based on the "common sense" reasoning that since all magnets are generally a given shape, then shape determines magentism.

An example from the same topic area of an item at the middle level of comprehension is the following:

Is it true that a magnetic compass which works on Earth would not work on Mars?

a. True -- because a magnet cannot work in a vacuum. [Content distraction]

b. True -- because Mars is too far from the North and South poles. [Intuition distraction]

c. False -- because all bodies of matter have a magnetic field. [Logical fallacy]

d. False -- because Mars has a magnetic field of its own. [Correct]

This problem requires the students to know about magnetism in the sense that they must recognize some parameters under which that natural force is functional. Distractors have been selected such that they test three several different error types of interest. The first distractor tests for content errors, since there is no real reason for it being selected except that with the mention of vacuums it sounds scientific (the presence or absence of a vacuum has no effect on magnetism). The second tests for intuition errors, since experience suggests the compass points north due to the presence of Earth's north pole, or that the focus of all magnetic fields is the polar region. The logical fallacy distractor is based on the assumption that all bodies of matter must have magnetic field since the Earth does.

An example of an item written to test the high level of comprehension is illustrated below.

If we were sitting in an army tank, is it true that a compass will still function properly and point to the north?

a. True -- Earth's magnetic field is so strong. [Intuition distraction]

b. True -- since the tank is made mainly of steel. [Logical fallacy]

c. False -- surrounded by steel, the compass will fail. [Correct]

d. False -- sound waves from the cannon will impair the compass's operation. [Content distraction]

The problem requires the student to know how magnets work, what is or is not magnetic, and that compasses function as a result of magnetism.

Clearly, several items of each level are needed in order to accurately diagnose a student's level of comprehension. Also, each item needs to cover as many as possible of the error types of interest in order to maximize the information gleaned for diagnostic purposes.

Experiment I

Method

Fifty three undergraduates in introductory biology classes at UCLA participated voluntarily as subjects in their regular classrooms. Nine items were written for each area of science to be tested (Photosynthesis, Magnetism, Energy in Closed Systems, and Gravity) representing three items at each level of comprehension, providing three opportunities to commit each error type. To these 36 items were added four more items in other areas of science. Test administration was handled as a conventional ad seriatim paper and pencil test.

Results and Discussion

The focus of the analyses conducted was to ascertain whether useful diagnostic information emerged from the use of either levels of comprehension or error types, or both. Protocols were scored for type of response made to each item -- correct, logical fallacy, intuition distraction, and content distraction. Table 1 presents the mean performance by level for each response type.

The results were encouraging. First, if students failed at any lower level, they were unable to succeed consistently at a high level of comprehension. For 89% of subjects in the sample, missing more than one item at any level of comprehension was accompanied by missing two or more items at the next higher level ($\chi^2 = 114.65$, $p < .001$). When considering each content area in isolation, the structure of the levels was maintained, although not to as great a degree. On average, 66% of the students responding in each content area failed at all levels higher than an initial error ($x^2 = 47.60$; $p < .001$). This confirmed the expectation that the levels of comprehension represent distinct levels of difficulty or complexity. The significance of this kind of information from a diagnostic test is that instructors could effectively gear their teaching to the precise level of comprehension at which a group, subgroup, or individual is presently operating.

Second, we examined the protocols for evidence of consistent error patterns within persons. We wanted to know whether the distinctions between error type would yield consistent diagnostic information concerning the type of errors which a given person was committing. Analysis of the protocols for the entire test showed no discernible pattern of error consistency. However, when the topic areas were examined separately, three distinct patterns emerged. First, content errors were the most consistent. On average, six persons were identified as making content errors within each area, a proportion significantly less than chance ($x^2 = 5.29$; $p < .05$). These were most probably guessing errors; people who guessed did so consistently. Their guesses were based on "test-wise" strategies in which scientific sounding, or stem-like responses were selected.

On the average, 21 subjects per content area consistently committed intuition or logical fallacy errors($x^2 = 4.43$; $p < .05$). Eighty four percent of these subjects committed one or the other type of errors consistently at all levels of comprehension subsequent to the first such commission ($x^2 = 35.52$; $p < .01$). Also, if either error type was committed at a low level of comprehension then one of the two errors was twice as likely to occur at the next higher level. The only problem encountered with the intuition and logical errors was that for 38% of subjects there was a tendency to shift from intuitive to logical fallacy errors, or vice versa, hence clouding the precise diagnosis of a certain error pattern. This problem was especially evident in the areas of magnetism and gravity. However, this proportion was not significant ($x^2 = 2.271$; $p > .05$), and so we expect that in future studies, with more items and subjects, such a problem will not prove detrimental.

Experiment II

A second study was conducted to investigate the degree to which the findings of Experiment I would generalize to other populations. Specifically, the questions of effectiveness concerning levels of comprehension and error types were addressed by administering a shortened version of the test to younger students of varied ability levels.

Method

Seventy six students from a private junior high school for the "gifted" (IQ measured above 145) in the west Los Angeles area participated, representing 38 seventh graders and 38 ninth graders. Sixty eight ninth-graders from a public junior high school in northeast Los Angeles also participated, representing three classrooms in which students were grouped according to common tracking procedures based on achievement. The latter subjects included 22 high, 20 middle, and 26 low achievers.

Subjects completed a short version of the same diagnostic science test developed for Experiment I. Twenty items with satisfactory psychometric properties were selected from among the 40 items from Experiment I; in each content area, two items were drawn from the high comprehension level, two from the middle level and one from the low level.

Results and Discussion

Two sets of analyses were conducted, one to assess consistency across level, and one to asses consistency across error type. Table 2 shows the mean performance in this test by subgrouping. The first analysis investigated the levels of comprehension ibedded in the items. For the 20-item test as a whole, results varied in consistency across the comprehension levels. Among higher ability students, the levels were relatively distinct and consistent; failures at a lower level of comprehension were accompanied by failures at higher level of comprehension for 67% of cases ($X^2 = 9.11$; $p < .01$). Among students with lower overall ability, the pattern was less clear, with reversals or sequence violations occuring in some 47% of cases ($X^2 = 20.28$; $p < .01$), although only 16% were cases in which total failure at lower levels was associated with total success at the higher level; ($X^2 = 3.44$; $p > .05$). The reasons for these contradictory results remain to be investigated fully. Several factors might explain the contradictions, including the reduction in number of test items, differences in mental ability or maturity, and differences in scholarship or classroom prepareness for test topics.

The next series of analyses concerned error patterns. While 45% of private school students committed no errors at more than one level, generally the high level ($X^2 = 15.82$; $p < .01$), 57% of the remaining students committed errors at multiple levels with consistent patterns for error type ($X^2 = 40.42$; $p < .01$). Unlike the patterns found for the college student sample, however, the most consistent patterns for the private school students emerged for logical fallacy errors. Public school students produced even less clear-cut results. Error types were committed in a nearly random manner for all but 34% of the students, with none of the three public school groups showing a significant proportion of subjects producing consistent error patterns when data were examined by topic area ($x^2 = 3.47$, n.s.).

Explanations for the lack of results in error patterns among both the public and the private school students center on two factors: the contrast in mental preparation with the college level group, and the reduced size of the test. Apparently, the more trained or "talented" the subjects are, the more consistency is found among the types of errors they commit. Intuitively this explanation makes sense if more preparation is associated with a lower probability of performing randomly, and hence a higher probability of systematic responses where one misconception might be reflected in several item responses. Because of the reduced test size, students had fewer opportunities to respond to items designed to re-measure the same concept and error types, hence less possibility for measuring consistent patterns.

Conclusions

The purpose of this study was to test the effectiveness of a prototype diagnostic science test. The test was designed to yield information concerning two dimensions: (1) the level at which students comprehend concepts within certain areas of science, and (2) the specific types of errors in reasoning which they systematically commit. We find the results encouraging in that they show the possibility of accurately measuring these two dimensions of reasoning (comprehension level and type of error) for purposes which should be of high utilitarian value to instructors.

Traditional tests which yield only a total and subscale test scores provide practitioners with less information which is readily employable for remedial purposes. The promise of diagnostic tests is that one can accurately assess not only the number of correct responses to test questions, but also the precise type of errors which are being committed and the level of comprehension at which a student could be characterized.

The results of the science reasoning test discussed in this paper suggest a variety of future research avenues in the area of diagnostic testing, studies which merit attention because of the value of diagnostic information. The inherent utility of adaptively structuring the student's path through a testing session is yet to be explored fully. This would allow the student to guided directly from item to item in a manner which matches subsequent items with patterns of distractor response seen in preceding items. Computerization of the science reasoning test is not difficult; the diagnostic interpretation available following an adaptive testing strategy could be enhanced relative to the conventional testing administered in the present study.

The expansion of suitable item pools in which item distractors follow a logical order and item complexity is layered seems essential. Further work in diagnostic evaluation of science understanding would entail writing item distractors which adhere to the remaining five error factors of Showalter (1974). However, not every item can be or need be accompanied by a distractor from every error factor. Such expansion should balance the additional error types across all item complexities. There will be a mandatory increase in test length, but all possible permutations of error types with one another cannot be brought together without excessive numbers of items. Thus further research in optimizing adaptive diagnostic test strategies is required.

Table 1

Probabilities by Response Type and Hem Complexity
Experiment I

Item Complexity	Response Type			
	Correct	Logical Fallacy	Intuition Distraction	Content Distraction
High	.58	.23	.17	.02
Medium	.78	.11	.08	.03
Low	.93	.03	.02	.02

Table 2

Probabilities by Response Type and Item Complexity Experiment II

Item Complexity:	Group:	Response Type Correct	 Logical Fallacy	 Intuition Distraction	 Content Distraction
High	Public Low	.10	.10	.12	.12
	Public Medium	.36	.27	.32	.05
	Public High	.25	.35	.13	.27
	Private 7	.37	.19	.35	.09
	Private 9	.48	.13	.32	.07
Medium	Public Low	.45	.27	.21	.07
	Public Medium	.57	.17	.17	.10
	Public High	.47	.22	.22	.08
	Private 7	.58	.24	.11	.07
	Private 9	.69	.15	.13	.03
Low	Public Low	.66	.26	.33	.23
	Public Medium	.27	.12	.09	.08
	Public High	.80	.05	.11	.04
	Private 7	.80	.06	.12	.02
	Private 9	.90	.02	.06	.02

REFERENCES

Billeh, V.Y. & Malik, M.H. Development and application of a test on understanding the nature of science. Science Education, 1977, 61, 559-571.

Cantu, L.L. & Herron, J.D. Concrete and formal piagetian stages and science concept attainment. Journal of Research in Science Teaching, 1978, 15, 135-143.

Cooley, W.W. & Klopfer, L. Test on understanding science. Princeton, New Jersey: Education Testing Science, 1961.

Dillashaw, F.G. & Okey, J.R. Test of the integrated science process skills for secondary science students. Science Education, 1980, 64, 601-608.

Dreyfus, A. & Jungwirth, E. A comparison of the prompting effect of out-of-school with that of in-school context on certain aspects of critical thinking. European Journal of Science Education, 1980, 2, 301-310 (a).

Dreyfus, A. & Jungwirth, E. Students' perception of the logical structure of curricular as compared with everyday context - study of critical thinking skills. Science Education, 1980, 64, 309-321 (b).

Gorodetsky, M. & Hoz, R. Use of concept profile analysis to identify difficulties in solving science problems. Science Education, 1980, 64, 671-678.

IEA, Test on understanding the nature of science. Stockholm: International Association for the Evaluation of Educational Achievement, 1969.

Johnstone, A.H. Diagnostic testing in science. In A. Levy & D. Neve (Eds.), Evaluation Roles in Education. London, Gordon & Breach, 1981.

Osborne, R. The analysis of student attainment of course objectives in physics. Journal of Research in Science Teaching, 1978, 15, 263-270.

Rodriguez, D.M. Notions of physical laws in childhood. Science Education, 1980, 64, 59-84.

Rubba, P.A. & Anderson, H.O. Development of an instrument to assess secondary school student's understanding of the nature of scientific knowledge. Science Education, 1970, 62, 449-453.

Showalter, V.M. What is united science education? Program objectives and scientific literacy. Prism II, 1974, 2.

TASK STRUCTURE DESIGN: BEYOND LINKAGE*

by

Eva L. Baker and Joan L. Herman

ABSTRACT

The role testing can play in ascertaining and improving the effects of educational programs and services is analyzed. Our point of view maintains that the connection between tests and instruction is best made integrally through an understanding of the design of learning tasks rather than through the use of techniques that attempt to join or to link the now-separate domains of instruction and testing. The context for task structures is described, and their use in developing instruction and tests is considered. The limitations of such an approach in practice are discussed and feasible approximations outlined. Finally, the research agenda in this area is broadly sketched.

Table of Contents

Page

Introduction 1

The Dimensions of the Problem 1

Testing in the Conceptual Framework of Science 2

Conceptual Framework for Design 5

Context: The World of Schools 6

Task Structures 9

1. Task Description 9
2. Content Limits 10
3. Generalization and Transfer 11
4. Discrimination/Performance Quality 15
5. Linguistic Features 19
6. Cognitive Complexity 20
7. Format 21

Introduction Implications of Task Structure Dimension 22

Applications of Task Structures 25

Theoretical and Applied Research Issues 28

References 30

INTRODUCTION

The purpose of this article is to analyze the role testing can play in ascertaining and improving the effects of educational programs and services. Our point of view maintains that the connection between tests and instruction is best made integrally through an understanding of the design of learning tasks rather than through the use of techniques that attempt to join or to link the now separate domains of instruction and testing.

The focus on the design requirements of learning tasks represents a fundamentally different perspective on the test/instruction issue. This perspective is theoretically grounded in its orientation deriving from research in learning, instruction, and cognitive processing, to name but a few areas; yet, it also has numerous potential implications for practice. The context for task structures will be described, and their use in developing instruction and tests will be considered. The limitations of such an approach in practice will be discussed and feasible approximations outlined. Finally, the research agenda in this area will be broadly sketched.

The Dimensions of the Problem

Public concern about the effectiveness of schools has led to a reliance on testing and test results that is unprecedented in recent educational history (Airasian & Madeus, this issue; Haertel & Calfee, this issue). Tests play prominent roles in certifying competency for high school graduation, in college admissions procedures, and in conveying through the publication of test results, the effectiveness of school district policies. These examples illustrate the practical and symbolic uses of tests. Test results are regarded as the "bottom line", and educators have devoted much attention to efforts to affect such scores, and thus graduate more students, place more in better colleges, and rank their district higher in test scores among other local school districts.

Since educators have accepted the validity of tests as outcome measures, they have fed the public's desire for accountability through testing, and have created a demon that needs continually to be satisfied. Yet, the goal of improving test scores is made extremely difficult by the ways in which schools are organized and staffed, by constraints on their resources, and by the trends in the society of which schools are only a part (Bank & Williams, 1982; Zucker, 1982).

Even if these factors were optimal, the problem of creating tests that are sensitive to the results of educational programs is a difficult proposition. Part of the difficulty stems from the way in which testing, as an enterprise and a research area, has developed and part from the existing scientific, conceptual framework of testing. In our view this framework is different from the conceptual framework of instructional design.

Testing in the Conceptual Framework of Science

A scientific and experimental orientation appears paramount in the psychometric view of testing. The practical uses of testing (to create a record of student performance) are subordinated in favor of the scientific use of tests (to detect individual differences or to determine effects of interventions). The focus on the latter purpose has led to a preference for the design and development of tests that best support these uses. For instance, in the Beginning Teacher Evaluation Study (BTES), items were not selected for inclusion on the dependent measure based on their fidelity to intended learning and instruction, but rather for their correlation in pilot data with the independent variables of interest (Filby and Dishaw, 1975). This procedure increases the probability that the test will find what the researchers are looking for,but does not insure that learning is assessed adequately--because the nature and definition of learning never emerges as a primary issue.

Test development from this scientific view follows a well-established, scholarly process. Tests are developed to assess a construct. Items for the test are selected or created, and data are collected on performance. Items are included, dropped, or revised in accordance with the fit the item data provide to the posited model. In this orientation, the psychometrician's job is to hypothesize a construct or trait that is assumed to be measured by a particular item set, and then to use that set to observe nature, to report reliably what exists, and to revise the test or to reformulate the construct to explain empirical facts, processes, and their relationships. The scientist-psychometrician's mind is active, but since his/her role is a descriptive one, that is, portraying how students respond to sets of items, the scientist remains outside of the action of instruction and passive with respect to creating "better" performance. The focus is on accuracy rather than on improvement. In fact, among the most serious errors a scientist can make is to perpetrate reactivity, where inadvertent effects are produced by the process of measurement itself. The measurer should make no ripples. This is one prong of passivity in world view.

A second prong of passivity derives from the notion of stability, a central thesis underlying much of science and measurement theory. The constructs to be measured are treated as stable and are described as traits or constellations of responses thought to persist in time. Classical psychometric and statistical theories have developed under the assumption that stability, regularity, and predictability are at the heart of scientific inquiries. A concommitant assumption, especially pertinent to this discussion, is that measured individual differences probably endure over intervention. Evidence for this point of view can be found in the literature on measuring change, where growth and measurement error are sometimes treated almost interchangeably (Harris, 1963). That the early uses of tests were for placement is no surprise, since labeling and grouping individuals in homogeneous clusters is a logical outgrowth of the belief in stability.

The conflict is most simply that education is an enterprise directed at producing change, yet our tests and the psychometric theory that generates and assures them are concerned with stability and description. Traditionally, measurement's role has been that of the objective, outside observer who analyzes its subject from afar. The serious integration of testing and instruction and the use of tests for instructional improvement, however, require a more active perspective that incorporates an inside view of the phenomena of interest, learning and instruction in the educational process.

Conceptual Framework for Design

As an alternative to the scientific perspective adopted by psychometricians, design methodologies reflect a different point of view. Design is a process that synthesizes practical and theoretically-grounded ideas to produce a procedure or product that changes the environment; it is an explicit problem-solving activity that generally includes notions of planning, creation, and fine-tuning. We are all familiar with design methodologies, from the most obvious aesthetic application (graphics, interior design), those which blend aesthetic and technical features (architecture), to those which emphasize the applications of scientific findings to particular problems (computer design, engineering of all sorts, medicine).

Most professional schools are committed to training at least some of their graduates in the design (as opposed to research) paradigm, although in education, as in other fields, design has somewhat less status than research activities. Because design creates things that must operate in reality rather than contend solely with the elegance of ideas, its place in the academic community is tenuous and probably survives because some design practitioners (doctors and lawyers) are perceived to be worth high levels of financial reward. Additionally, although the outputs of design may include scientifically-based processes, it is undeniable that art or craft is also demanded. Thus, design work gets labelled "atheoretical," and its status denigrated. Even status aside, because we know less well how to help people become talented designers than competent researchers, design aspects in education are often neglected.

In contrast to the researcher, the task of the designer is not descriptive. Rather, from the outset, the designer's task is to improve upon present practice. Teachers in our public schools are design practitioners, however informally they accept that role Their task is to combine information from a variety of theoretical approaches with practical wisdom to affect the quality of education. Change is the goal of a designer and thus that orientation would seem to serve education well, and particularly the problem of connecting tests and instruction.

Even if one were to minimize the effects of these two different frameworks on the predisposition for action or reflection, or perhaps admit that there is a continuum rather than two mutually exclusive perspectives, the serious problems remain in connecting testing to instruction and promoting school-based change. These problems inhere in the realities of school operations and in the nature of effective instruction.

Context: The World of Schools

Because psychometrics is a scholarly pursuit, most of its work is conducted in settings remote from current public school experience. Psychometric researchers don't often go to schools, and when they do, they usually don't focus on instructional issues. Because of this lack of familiarity with the lives of people in schools, it is not surprising that some misperceptions seem to have occurred.

One set of misperceptions involves teachers and what it means to teach something. Another concern is the extent to which the process of curriculum development, adoption, and implementation occurs and can be counted upon to provide a common context for school events in different schools (Sirotnik, 1981). For instance, the irrationality of the system as it most frequently operates causes only an occasional lament on the part of psychometricians. Consider that school curricula, and therefore, most of formalized content, are developed outside of the schools and marketed by text publishers. The same is true for tests used in schools. Unfortunately, the coordination between test and curriculum development is nil. Thus, at the most basic level, we can show that content differs in tests and in texts (See Floden, Porter, Schmidt & Freeman 1980) and that at the grossest exposure level, students cannot be expected to perform with regard to content they have not seen. A recent study comparing district curriculum objectives in both language and mathematics with the state assessment and various standardized tests illustrates this irrationality. Some tests included only 25% of the district curriculum, and, in some, almost 50% of the test was not covered by the curriculum (Cabello, 1982).

But attention to curricular match demonstrates only very global concern for the relationship between testing and instruction. Such a concern assumes that formal curriculum is, in fact, implemented and that simple exposure to content is sufficient for students to learn; both assumptions are unfounded. In seriously coordinating testing and instruction, one immediately is confronted with the complexity of actual instruction, including, for instance, how learning takes place, the range of content presented, the context in which a set of skills should apply, and how teachers augment or circumvent existing text material to facilitate learning. One might be tempted to ignore these confounded variables and to try to bring order to the system gradually by focusing on and controlling only one set of variables at a time. Yet, when one focuses on one aspect, say, content, and attempts to match tests and curricula on that basis, one necessarily ignores other important considerations that contribute to the irrationality of the present system. The status quo, as a result, is inadvertently perpetuated . Unless integrated alternatives are pursued deliberately, the status quo with its irrational base will continue to be our only option.

We propose, therefore, to use design methodology to create an improved system for education and learning. This approach concentrates not on linking extant curricula with extant tests, but instead on the complete design of an entire system; in fact, a redesign, so that the entire system makes sense. Instead of studying testing properties in terms of existing rules of order, we propose to focus on the learning tasks of students. The system starts with the nature and definition of what is to be learned--the task structure; the characteristics of instruction and of testing then follow naturally and rationally. Our proposition is that by designing task structures, we provide a model of the features that learning should exhibit. The model causes the requirements for testing and teaching to converge since they share, by definition, critical features. Linkage becomes redundant.

Task Structures

The task structure approach is based upon the design of the learning tasks desired of the learner. This structure is specified by a series of rules or examples; the final object is to present as clearly as possible the expected set of skills with regard to specific content. Task structures integrate critical ideas in education by reserving a place for them in the structure itself. For example, the range of content over which the learner's skill is to generalize, the manner in which transfer is treated, the behavioral formats for exhibiting performance, the level of cognitive operations required, and the complexity of language are all explicitly treated in the task structure.

We will treat the features of a task structure in turn and present definitions and descriptions, theoretical connections, and examples.

Task Description

The first element in a task structure is the general description of the task. This statement may be thought as equivalent to the statement of objective in objective-referenced tests or outcome statements used to guide the development of criterion-referenced tests. Its purpose is simply to direct and circumscribe attention to a general area of content, such as geometric proofs, and to focus on the type of skill needed, such as to solve problems or to demonstrate procedures. Such statements often have served as the sole descriptor for developing criterion-referenced tests. However, in a task structure, the statement serves principally as a convenience, as a way to get into or to approach the more taxing endeavor of describing the learning requirements of a particular task.

Content Limits

Establishing the content limit of the task is an initial problem. These limits are intended to circumscribe clearly the substance or content upon which the learner is supposed to operate. While the task description describes content in very general terms, content limits make more specific the particular elements of content to be included, and thereby help to make explicit requirement for instructional exposure and opportunity to learn. Two common approaches to content limits have been advanced: definition by curricu[illegible], and definition by agreement.

The first approach responds to the real world of school exigencies and suggests that content be defined by reference to extant curricular material (Baker, 1974), such as specifying permissible test content for a reading comprehension objective to include non-fiction selections occurring in a particular 9th grade district adopted textbook. In this instance, the content is selected based simply upon the rule of potential exposure, and while opportunity to learn is an obvious criterion, the selection of substantive material has been left to the indeterminate judgment of textbook writers and publishers. Unfortunately, analyses have shown that systematic structure and features that contribute explicitly to learning are often absent in commonly used texts (Quellmalz, Herman & Snidman, 1977; Herman, Hanelin & Cone, 1977). The reading selections at a particular grade level, for instance, do not appear to follow inferrable rules of progressive linguistic, semantic, or syntactic complexity.

A second common approach to content limits involves developing an agreed upon set of boundaries which is disseminated to both teachers and test writers. The benefits of such agreement is that the probability of a fit between instruction and testing is increased simply by communication, permitting an instructionally-focused outcome system (Popham, 1981). The criticism of such an approach rests upon the arbitrariness of the content boundaries selected (Why four line paragraphs instead of five or three line in a reading comprehension task, for example). This charge of arbitrariness has been countered by appeals to the wisdom of reliance on "human judgment."

We suggest that arbitrariness of content limits may be mitigated through reliance upon relatively strong theoretical or empirical knowledge about learning (as well as on human judgment) to decide what content limits are sensible. This approach substantially supplements the rationalization of content based upon probable exposure to extant curricula, e.g., a 9th grade textbook, and the human judgment defense, and permits each element of content limits to respond to potential issues in learning. We propose using two well documented constructs related to research on learning to define content limits: 1) generalization and transfer; 2) quality of discrimination/performance.

Generalization and Transfer. The concepts of generalization and transfer help alleviate the anathema "teaching to the test". Psychometricians, perhaps because of their frequent concentration on individual items rather than on concretely related item sets, seem to worry about this problem a good deal. Not acknowledged is the fact that "teaching to the test" can occur in two forms: teaching the exact items that appear on the test and teaching the class of content, by sampling, that the test is designed to measure. The former appeals to rote learning and is usually of limited educational value; the latter demands that we have a solid notion of what learning is intended and that we focus on significant higher-level tasks. The content limits should create this notion of the class of substance to which the test behavior is meant to apply and generalize, echoing the general idea of domain-referenced testing (Hively, Patterson, & Page, 1968).

One question here is to what topic or topics is the behavior supposed to generalize? The answer to that question depends generally on practical matters, such as the length of instructional time available for the task, or more directly, to the level of specificity at which the task is to be learned, as well as inter-relationships among potential topics. For instance, a learning task related to conjugating and applying "-ar" verbs in Spanish might be defined with the expectation that the learning should generalize to all such verbs. Alternatively, the task could be expanded to all regular "-ar, -er, and -ir" verbs, requiring a longer allocation of instruction. The conjugation procedure for these Spanish verbs is essentially the same; only details change and change in predictable ways. Research on generalization and transfer would suggest that teaching students the critical features of such conjugation and providing related practice would be sufficient for them to apply the rules to any new examples of regular Spanish verbs, and thus would support the more inclusive task structure. In contrast, the coherence and research support would be more problemmatic if the content for a task in Spanish included verb conjugations, pronoun number, and sentence expansion. The probability that transfer would occur is low, precisely because the information and concepts necessary for success vary. One decides on what level of generalization and transfer one can achieve partly by relying on unfettered human judgment, but partly based upon theoretical or empirical evidence about how knowledge in a particular subject matter area is connected (Geeslin & Shavelson, 1975).

Specifying the topics over which performance is to generalize serves an additional purpose: if transfer between topics is not explicitly taught, there is little reason to believe transfer will occur (Silberman, 1964). The definition of task structures in written composition, for example, relates to the low transfer of writing ability between topics, such as "My Best Friend" and "Credit Card Use in the United States" (Quellmalz, Capell, & Chou, 1982). Students need to learn that the same strategies and skills are applicable in both cases. The explication of the range of topics over which performance is expected or desired to generalize, when communicated to those responsible for instruction, then, can itself facilitate transfer.

Since tests take considerable time from instruction, they probably should be perceived seriously and reserved for those goals that incorporate generalization and transfer rather than for memorization of specific content. Desired levels of generalization and transfer among topics should be specified based upon the level of effort in instruction and theories of content relationships derived from analyses of the discipline (in addition to the more common curricular and consensual bases described above).

A second area to be considered under the category of generalization and transfer relates to the form in which information is presented to the learner. Form is not item format, such as passage length or number of distractors, but rather the substantive features of the task, other than topic, over which the learner is expected to generalize. For example, "triangles" is one topic in the task of learning to discriminate confusable geometric figures. To ascertain success, the learner can be asked to discriminate the correct answer presented in a single form, e.g., your standard, equilateral triangle. However, if one wants to assure an understanding of a triangle that is somewhat more robust, one would provide students with correct answers that include acute or right triangles, and perhaps triangles whose vertex is not perpendicular to the margins of the page. Students might be asked to find the triangle when other salient perceptual cues, such as size and color, might interfere. Here the issue is clearly "over what cases does the learner recognize a triangle?"

Form and type of information can be illustrated in non-perceptual concept learning as well. What class of information will the learner be expected to have acquired in order to attempt the task? For example, in written composition, a question may be posed about whether the learner has sufficient knowledge about a topic to write about it. How is that information to be provided? How complex will be the form in which the information is presented? Will it be a list that the learner simply has to transpose into prose? Will he/she be expected to infer meaning from embedded and subordinated information? As a second example, consider the task of learning to identify the main idea of a prose passage, a common enough objective of reading instruction. What type of passage will be presented? Will it be one in which there is a single clear main idea? Or one in which two partially developed ideas compete with a third "main" idea for dominance? The differences in the task intellectually should be clear, and the different requirements for instruction are probably obvious.[1]

We are recommending, then, that issues of generalization and transfer be incorporated in the content limits section of a task structure specifically to address the topics over which the response is supposed to generalize and information or presentation forms over which the response should transfer. Both of these areas are to supplement the simple notion of opportunity to learn, defined either as content in required texts (Floden, et al, 1980), or time on task, i.e., time nominally allocated to particular topic (Denham & Lieberman, 1980). It is our belief that these instances are too global to relate productively to learning tasks. In addition, we believe that attending to generalization and transfer strengthens human judgment because theoretical and empirical bases are used for content selection rather than more vague appeals to authority.

Discrimination/Performance Quality. A second, general area within the content limits section of task structures focuses on the standard of performance expected of the learner. It is at this point where claims of "educational excellence" are based by defining the required quality of response. However, performance quality should not be confused with common versions of performance standards (Mager, 1961; Popham & Baker, 1968; Anderson & Faust, 1973), all generated during the behavioral objectives era of the sixties and seventies. Performance quality relates to non-quantitative features of responses that illustrate the level of refinement of the response. Because student response options fall into two major categories. Let us illustrate this principle in both selected and constructed responses.

---

[1] Directly related to the issue of form of information is the linguistic features of the text of the test items. However, we will separate that discussion into a later component of task structure because the theory which supports such analyses differs from the cognitive research base of the present section.

In a selected response task such as the triangle discrimination task described above, the *a priori* difficulty of the task depends upon not only the range of correct answers the learner has to identify, e.g., isosceles and acute triangles, but the fineness of discrimination required to make that identification from distractors. Very little refinement would be required to select from distractors that consisted of those in panels a or b in Figure 1. Consider, however, if the distractors consisted of those shown in panel c. An analysis of these latter response options should demonstrate that the item requires relatively fine discriminations and exhibits higher *a priori* difficulty.

Insert Figure 1 About Here

The analysis also should clearly demonstrate that choice of distractor provides diagnostic information about the class of mistake the student is making. To select the first option in panel c, the student would have to believe that open as well as closed, three-sided, straight-lined figures met the definition of triangle. For the

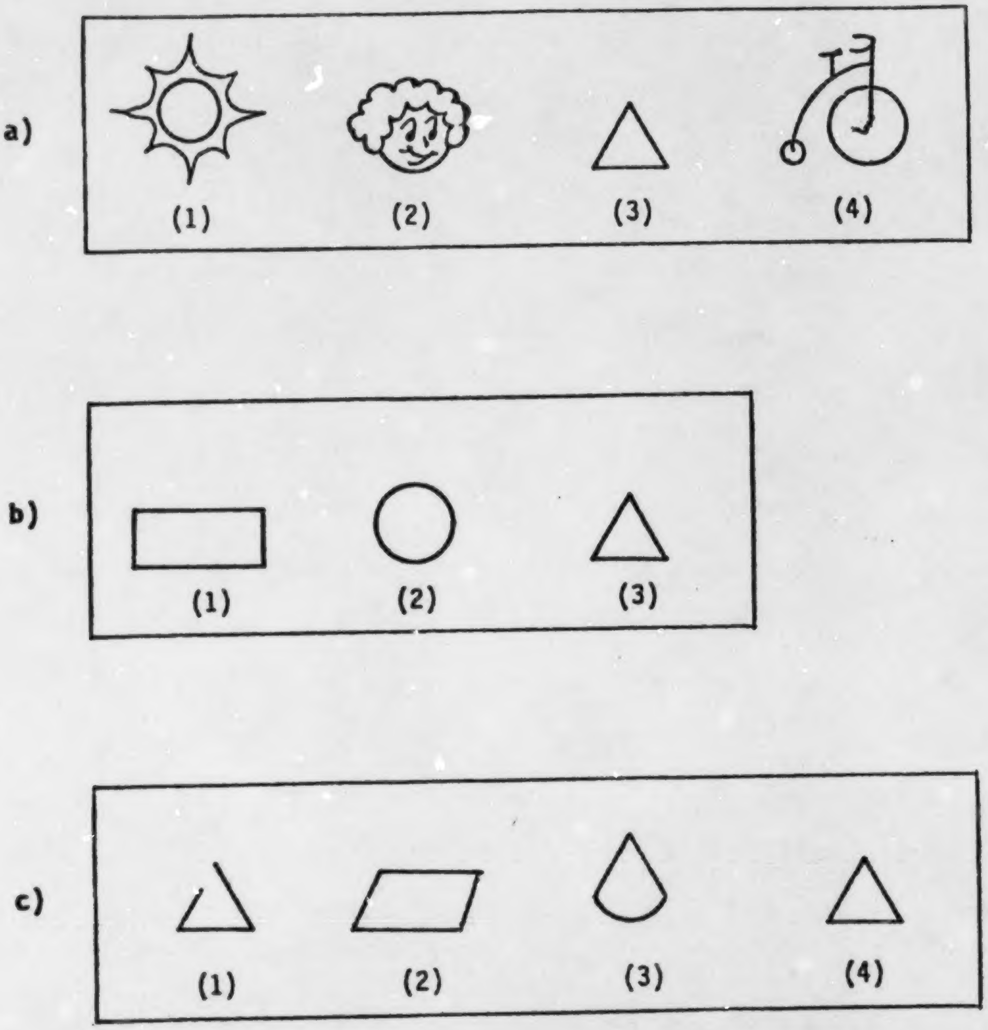

Figure 1

Alternative Distractor Options for a Triangle Discrimination Task second option, the student would have neglected the three-sided aspect of the concept. In option three, the student would have overlooked the requirement of straight sides. In each of these instances the provision of diagnostic information is preplanned.

In contrast, much of the extant literature in diagnostic testing (see Brown & Burton, 1978; Tatsuoka & Tatsuoka, 1980) lacks such *a priori* design of distractors to yield explicit diagnostic information; instead, sources of error are inferred from item response patterns. Explicitly including rules for the creation of wrong answer domains in the content limits can significantly increase diagnostic power, and to the extent that such rules incorporate research on concept learning (such as Tiemann & Markle, 1973; Tennyson, Wooley, & Merrill, 1972), the diagnostic quality will be more refined. Where concept learning is not the focus, content limits for multiple-choice items may be generated specifically to deal with aspects of the task that may have been underlearned as well as those aspects that may have been mislearned.

In the case of constructed response, where no distractors are provided for the learner, the content limits should account for the explicit standards that will be used to judge the quality of the student effort. These standards, or criteria, are applied to student products to assess the extent to which products, such as essay answers to science questions, or English compositions, exhibit desired features. Such decisions can be reached through holistic approaches, where the overall value of the paper is judged by internal standards, or by analytic methods where particular aspects of student production, such as style, coherence or grammar, are separately considered. In either the holistic or analytic approach, the response may be judged according to a check list (where the paper, or the style, is either satisfactory or not), or through the application of a rating scale (where points from 6 to 1 depend upon the quality of student performance.)

It should be clear to see that less well explicated scoring systems, i.e., holistic, rely more on undifferentiated human judgment and experience, whereas explicated standards, such as analytic approaches with logically anchored rating scales, provide much more information about student performance. This additional information is desirable in task structures for it directly implies the type of instructional tasks the learner is expected to encounter as well as the remedies that may be necessary to address inadequate performance. Explicated standards for judging criterion responses, thus, are an important component for teaching and testing.

With the specification of content limits, performance quality is measured by design, either inherent in the level of discrimination required in selected responses or by the explicit statement of criteria in production responses. Difficulty emerges directly from the task structure design and is a function of task complexity and fineness of required discrimination rather than created emprically by proportions of people who succeed at an item. This conceptual design of difficulty may help break the tautology that exists between empirical "item difficulty" and assessment of the effects of instruction. Such an approach also allows one, by reviewing wrong answer choices, for instance, to determine when partial learning has occurred and where remediation is needed.

Linguistic Features

Linguistic features are another important aspect of task structures, but their role in test and task design has been treated in generally disjointed fashion. Level of difficulty has been assessed by various readability formulae which take into account the difficulty level of words (inferred from developmental or frequency measures), and sometimes the complexity of syntax (Duffy, 1981). Yet more complex linguistic structures play a role in tasks that either present verbal material as stimuli, including verbally stated alternative responses, or include rating systems based on verbal products by the respondent (Duffy, Curran, & Sass, 1982). Particularly when non-native English speakers are assessed, the variation in performance created by apparently casual linguistic options may be great. Bauman (1982), for example, found that problem types identified through linguistic analysis posed serious difficulties for readers--problems that were not directly related to the construct being assessed.

Systematic attention to the linguistic components of tasks may permit more accurate assessments of true performance levels. Measures of linguistic complexity need to be created that are appropriate for both long and short verbal passages and which include some notion of deviation from semantic and syntactic experiences of the respondents. For instance, some difference score may be obtained depending on the compatibility of the sentence patterns with the native language, or the root of more difficult words and the native language. To the extent that language proficiency is not an inherent feature of the task of interest, then effort should be made to purge verbal materials of unnecessary complexity.

Cognitive Complexity

Another critical feature of a task structure is the cognitive complexity of the task. Simply stated, cognitive complexity is the intellectual "level" apart from content, at which the learner is expected to perform. These levels have been taxonomized by Bloom and his associates (1956) to include six categories, Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Presumably, each of these categories refers to cognitive processes that are successively increasingly complex as well as dependent upon prior levels. A slightly different structure has been posited by Gagne (1975) where essentially stimulus-response learning, multiple discrimination, concept learning and problem solving form the major dimensions of intellectual skills. Simplifications of these schemes have been found in the cognitive literature (Quellmalz, 1982) where principal distinctions have been made for tasks whose purpose is the storage, association, and retrieval of information contrasted with tasks requiring processing of information, including subordination, reconfiguration, and other adaptive processes.

The task structure must clearly provide an indication of the intended cognitive complexity of the task. This ascription will relate to two features of the content limits already described, generalization and required performance quality. First, complexity is a function of the degree of expected generalization and transfer, and in the nature of the required performance, e.g., the number of cues provided, the amount of information that must be subordinated should be reflected in this area. Second, complexity is also a function of the performance quality that is demanded. A good example might be a problem solving task involving the correction of a operating defect on a jet aircraft. Perhaps no transfer is necessary, for it is the F-14 and only the F-14 that is of interest. However, because of the enormous inherent complexity of the circuitry of various systems, the task requires within it a high level of discrimination and therefore, has high cognitive complexity.

Format

Another related feature of the task is the format in which the assessment is made. Format includes both the descriptive mode in which the task is presented, e.g., print, graphics, video recording, and the form in which the response is desired, e.g., multiple choice with four response options, written composition, and so on. Obviously the format relates both to the practical matter of presenting and obtaining task related information as well as to the requirement to incorporate specifications of task structure identified in content limits, in linguistic features, and in cognitive complexity. It is possible, for instance, that format is truly an unimportant issue, and that fact is demonstrated by the expectation that students will be able to demonstrate task mastery in one of many formats or in all of a number of formats.

The extent to which format dependence has taken over from optimal learning requirements of tasks is documented by the attention the general education system directs to test wiseness. Here the format of the test is regarded as separate from, and sometimes equal to (in importance), the content and intellectual skill demanded by the task itself. Including format as a particular dimension of the task structure allows for the rational review of the role of format and its relative importance or subordination to issues of content mastery. In addition, the internal consistency of descriptions about generalization of content performance quality and cognitive complexity can be assessed in reviewing the format(s) projected for task demonstration.

Instructional Implications of Task Structure Dimensions

The premise of this paper is that attention to task structure dimensions outlined above provides a common focus for and defines the structure of assessment and instructional systems. In this section, we propose to identify the aspects of the task structure that inexorably lead to instructional decisions. The problem in relating instruction and assessment changes dramatically. Instead of dealing with the amount and degree of overlap between activities and artifacts, one focuses on the degree of implementation of the task structure itself, a far different task intellectually, and with the potential, at least, for greater satisfaction.

In dealing with instruction, let us exclude from our discussion issues related to affective, motivational or social learning paradigms and focus, for purposes of our analyses, on the cognitive and behavioral tasks of learning and teaching. Clearly, based on the literature in instruction (Bower & Hilgard, 1982; Gagne & Briggs, 1981; Traub, 1966), a critical issue is the extent to which students have been exposed to a particular task and in fact have had the chance to practice it under conditions implied by the content levels, i.e., with both particular and generalized examples, and at the level of performance quality (such as discrimination), implied by the task. However, opportunity to practice criterion behavior is necessary but may not be sufficient for less able students and more complex tasks.

If criterion behavior is too complex to be acquired by repeated rehearsals, what should be done first, what component skills must be acquired? Unlike many statements of objectives, the subordinate components of instructional tasks are inherent in the task structure itself. The identification of features over which the performance is expected to transfer specifies a set of experiences for the student. For example, if the task structure is to be able to analyze particular propaganda devices in advertisements, news articles, editorials, and verbal appeals, then students would need practice with all specified media as well as instruction and practice with each specified device. The indicated embeddedness and subtlety of propaganda use would similarly suggest the successive range of difficulty that would be appropriate for instruction. In other words, inherent in the task structure is a plan for successive approximation of the end desired learning tasks, where individual components are practiced and then combined in increasingly complex sets.

The nature of instructional tasks also follows from the specification of content limits for performance quality: the classes of concepts included in the distractors, or the criteria by which the ultimate student product is to be judged. For instance, in the triangle discrimination task described above, instruction would need to take clear and differentiated account of the attributes of the triangle of interest: it is a geometric figure; it has straight lines; it is closed; it is three sided. The order in which these are treated or the motivational context in which these attributes are introduced make little difference to this analysis. The implementation issue is the extent and degree to which these attributes are treated, i.e. the extent to which instruction and practice deal with each attribute, singly and/or in combination, which represents a significantly more refined view of opportunity to learn.

Similarly, in constructed responses, if a learner's writing is to be judged on his/her use of coherent sentences in a paragraph and the choice of development used in the paragraph, then the instruction must, in a differentiated way, treat these options. Again, the context in which instruction occurs or the instructional approach, is not of first concern; matters of presentation style, sequence, etc., are not the primary focus because valid differences cannot be discerned in the absence of specified treatments, treatments which are directly relevant to and derived from the desired learning. Once more, the issue becomes whether the elements of the task structure can be found in the instructional provisions for the students. It is an implementation problem, looking at frequency and intensity, rather than a problem of determining overlap.

Metastructures for approaching these individual instructional components of the task structure depend on the educational philosophy and instructional style preference of the teacher. Direct instruction (Rosenshine, 1982) and task analyses (Gagne, 1977) approaches to teaching may be appropriate. On the other hand, a less directive, more inquiry-oriented approach may be preferred. Most important to note, however, is that the action changes from attention to the process of instruction, or how instruction occurs, to the substance of instruction and the modelling of the structure of task itself.[2] Our belief is that the way to outcomes is far easier and of secondary importance if the quality of outcome desired is sophisticated and well described. Targetted instruction under whatever approach, will likely be more effective than more diffuse attempts. You have to teach "it" if "it" is going to be learned.

Applications of Task Structures

Since it is obvious that the rhetoric of design and change is insufficient itself to create the conditions for implementation in education, what is the likelihood that such an approach is practical at all? Organizations responsible for implementing educational practices like public schools, are often not change-oriented themselves. They would rather adopt the surface appearance of change and innovation (Pincus, 1975) than to undergo the dislocation that real change implies.

Having laid out our ideas on task structures and the promise they hold for making the educational process more rational, fair, and instructionally effective, let us consider their possibilities in practice. Or have we, like the academic friends we've criticized, proposed an ivory tower system that will not survive the test of reality?

First, let us consider a serious distortion of the ideas we espouse: minimum competency testing. Essential in this movement is

---

[2] We are indebted to Wells Hively for this part of the analysis.

the idea that schools should be responsible for assuring the acquisition of particular skills--learning tasks--and that these skills should be the subject of both instruction and testing. Yet, in practice, the target skills do not truly reflect school and teachers' main goals, and the natural linkage of instruction and testing within the system has not often occurred. Insufficient technical expertise, often volatile political environments, and high stakes have combined to produce more rather than less irrationality: ninth grade--or lower--skills many represent essentially a new one-year remedial curriculum masquerading as the minimum competency for high school graduation. Such may be the fate of most top-down change mandates that attempt to solve complex educational problems with simplistic solutions that are insensitive to local context.

Our experience, however, indicates that more positive outcomes are possible, and that approximations of our learning task approach are feasible in practice. Below we allude to two approaches we have used to implement task structures. The two examples vary in the local motivation for change and the source of educational goals--or the learning tasks to be accomplished. The examples illustrate a "minimal" and "maximum" attempt at change.

With a minimalist view, one school district attempted to solve a common district problem, "Raise those test scores." Learning tasks were directly inferred from the actual content of the tests in question, i.e. the task structures were defined to parallel state assessment test content. District curricula were analyzed to determine the extent to which instruction and practice were provided for each learning task. Little direct and explicit overlap was found and supplementary practice exercises and cues for instruction were developed to fill in the gaps. Additionally, test performance was analyzed school-by-school within the district and school specific instructional prescriptions were created. School-wide strategies and explicit instructional guidance and materials for teachers and students were designed. The entire effort was initiated centrally and received strong district leadership and subsequent principal support. While some might question the validity and value of such "teaching to test" activities, the effort was directed at instructional improvement, based on the goals measured by the test, and served the practical needs of the subject school district.

More comprehensive change efforts have been conducted in other local contexts, using a more grass roots approach. Several change efforts conducted by the UCLA Center for the Study of Evaluation have used a multiprong curriculum-assessment-staff development strategy. In these instances, teachers have been trained in the task-structure approach to integrating instruction and testing and in sound test development techniques. Teachers then play the active role, with some technical assistance, in defining critical learning tasks for their subject area, in explicating the dimensions of each task and in constructing suitable test items. The resultant tests are subsequently used to diagnose individual, class, school, and district needs, and to monitor student achievement. Model instructional approaches and teaching lessons for the target learning tasks also support the process.

More comprehensive implementation of task structures is possible. Applications in emerging technology and in the private sector in highly technical training environments represent two potential opportunities. In both cases, the incentive for high quality training may be possible to an extent not present in public school education. In the second case, there are controls on the selection of the group to receive education, either because they are hired or otherwise screened, and teaching conditions and student motivation are more tractable. The use of technology is a seductive arena not only because the personal and idiosyncratic mediation of instruction by teachers will be avoided, but also because of the possibilities for closer monitoring and immediate feedback with refined branching and remediation options.

Theoretical and Applied Research Issues

The role of theory in research on task structures is obvious. However, the theory of interest is not psychometric theory, but rather propositions that grow from perspectives in cognitive and behavioral learning in the field of psychology, in psycholinguistics, and in contrastive linguistics. A practical issue relates, once more, to the level of generality necessary and the inherent relationships among features of the task structures. For example, can one have relatively simple content and require sophisticated cognitive processes? The answer on a single instance level is "of course", but how general is that answer?. What is the relationship between language complexity, cognitive processes, and transfer and generalization of content? How circumscribed or broad can a task structure be; that is, what are the limits or optimal levels of generalization?. These and other more provocative questions need exploration as well as testing in alternative contexts and degrees of implementation. At any rate, what we hope will happen is that those with psychometric skills and those whose expertise is in the areas of learning and instruction will meet intellectually and jointly continue the task of focusing educational productivity on learning tasks.

REFERENCES

Anderson, R.C. & Faust, G.W. *Educational Psychology--The Science of Instruction & Learning*. New York: Harper and Row, 1973.

Baker, E.L. Beyond objectives: Domain referenced tests for evaluation and instructional improvement. *Educational Technology*, 1974, *14*, 16-21.

Bauman, J. *Linguistic Structure and the Validity of Reading Comprehension Tests (Final Report to National Institute of Education)*. Washington, D.C.: Center for Applied Linguistics, May 1982.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (Eds.) *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay, 1956.

Bower, G.H., & Hilgard, E.R. *Theories of Learning*. Englewood Cliffs, New Jersey: Prentice-Hall, 1981.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 1978, *2*, 155-192.

Cabello, B. Analysis of district curriculum, CAP objectives and 3 standardized tests. Los Angeles, California: Center for the Study of Evaluation, 1982.

Denham, C., & Lieberman, A. (Eds.) *Time to learn*. Washington, D.C.: National Institute of Education, May 1980.

Duffy, T.M. *Readability formulas: What is the use?* San Diego, California: U.S. Navy Personnel Research and Development Center, November 1981.

Duffy, T.M., Curran, T.E., & Sass, D. *Document design for technical job tasks: An evaluation*. San Diego, California: U.S. Navy Personnel Research and Development Center, April 1982.

Filby, N., & Dishaw, M. Developments and refinements of reading and mathematics tests for grades 2 and 5. Technical Report III-1, Beginning Teacher Evaluation Study, San Francisco, California: Far West laboratory for Educational Research and Development, 1975.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. Don't they all measure the same thing? Consequences of standardized test selection. In E.L. Baker and E.S. Quellmalz (Eds.), *Educational testing and evaluation*. Beverly Hills, California: SAGE Publications, 1980, 109-120.

Gagne, R.M. Analyses of lectures. In L. Briggs (Ed.) *Instructional design: Principles and applications*. Englewood Cliffs, New Jersey: Educational Technical Publications, 1977.

Gagne, R.M., & Briggs, L.J. *Principals of Instructional Design*. New York: Holt, Rinehart, and Winston, 1974.

Geeslin, W.E., & Shavelson, R.J. An exploratory analysis of the representation of mathtmatical structure in students' cognitive structures. *American Educational Research Journal*, Winter 1975, *12*(1), 21-39.

Harris, C.W. (Ed.) *Problems in measuring change*. Madison, Wisconsin: University of Wisconsin Press, 1963.

Herman, J., Hanelin, S., & Cone, R. Instrumentation for the Early Childhood Education Program. Paper presented at the annual meeting of the American Educational Research Association. New York, 1977.

Hively, W., Patterson, H., & Page, S. A "universe defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*(4), 275-290.

Mager, R.F. *Preparing instructional objectives*. Palo Alto, California: Fearon Publishers, 1961.

Pincus, J. Incentives for innovation in the public schools. *Review of Educational Research*, 1975, *45*(1).

Popham. W.J., & Baker, E.L. *Rules for the development of instructional products*. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development (SWRL), 1968.

Popham, W.J. *Modern educational measurement*. Englewood Cliffs, New Jersey:Prentice-Hall, 1981, p. 211.

Quellmalz, E. Cognitive models for linking testing and evaluation. Los Angeles, California: Center for the Study of Evaluation, 1982.

Quellmalz, E., Capell, F., & Chou, C. Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, 1982, *19*, 241-258.

Quellmalz, E., Herman, J., & Snidman, N. *Toward competency-based reading systems*. Paper presented at the Annual Meeting of the American Educational Research Association. New York, 1977.

Rosenshine, B. *The master teacher and master developer*. Paper presented at the Annual Meeting of the American Educational Research Association. New York, 1982.

Silberman, H.F. *Experimental analysis of a beginning reading skill.* Santa Monica, California: Systems Development Corporation, 1964.

Sirotnik, K.A. *The contextual correlates of the relative expenditures of classroom time on instruction and behavior: An exploratory study of secondary schools and classes* (A Study of Schooling Tech. Rep. No. 26). Los Angeles, California: University of California, Laboratory in School and Community Education, 1981.

Tatsuoka, K., & Tatsuoka, M.M. *Detection of aberrant response patterns and their effects on dimensionality* (Research Rep. 80-4). Urbana, Illinois: University of Illinois, Computer-Based Education Research Laboratory, 1980.

Tennyson, R.D., Wooley, F.R., & Merrill, M.D. Exemplar and non-exemplar variables which produce correct classification errors. *Journal of Educational Psychology*, 1972, *63*, 144-152.

Tiemman, P., & Markle, S.M. Remodeling a model: An elaborated hierarchy of types of learning. *Educational Psychologist*, 1973, *10*, 147-158.

Traub, R.E. Importance of problem heterogeneity to programmed instruction. *Journal of Educational Psychology*, 1966, *57*, 54-60.

Williams, R., & Bank, A. Use of data to improve instruction in local school districts: Problems and possibilities. In C. Aslanian (Ed.), *Improving Educational Methods: Impact on Policy*. Beverly Hills, California: SAGE Publications, 1981.

Zucker, L.G. Institutional structure and organizational processes: The role of evaluation units in schools. In A. Bank & R. Williams (Eds.), Evaluation in school districts: Organizational perspectives, *CSE Monograph Series in Evaluation*, 1981, *10*, 69-90.